

A genome-scale CRISPRi perturbation atlas of human induced pluripotent stem cells

Received: 5 September 2025

Accepted: 29 May 2026

Published online: 01 July 2026

 Check for updates

Sami Nourreddine^{1,16}, Yesh Doctor^{1,16}, Amir Dailamy², Yi-Hung Lee¹, Jan N. Hansen³, Rebecca Chinn^{1,4}, Antoine Forget^{5,6}, Benjamin Polacco^{5,6,7}, Monita Muralidharan^{5,6}, Alina Sigaeva⁸, Sushant Sunder¹, Emily Pan¹, Jiahao Gao^{9,10}, Jake Y. Chen¹¹, Timothy Clark¹², Jillian Parker⁹, Kirsten Obernier^{5,6}, Christian Metallo^{1,4}, Trey Ideker^{1,9,10}, Emma Lundberg^{3,8,13,14}, Nevan Krogan^{5,6,7,15} & Prashant Mali¹✉

Comprehensively mapping the relationship between genotype and phenotype offers essential insights into how a cell's state arises from its genetic components. Toward this goal, we generated an expressed genome-scale CRISPRi perturbation cell atlas in KOLF2.1J human induced pluripotent stem cells, mapping transcriptional phenotypes associated with 11,692 perturbed genes across >2.5 million single cells. Using correlations among perturbed phenotypes, we created a cell map of the pluripotent state, demonstrating rich recapitulation of functionally related protein complexes. We then explored the atlas to uncover metabolic factor ZBTB41 and pluripotency regulator RNF7, validating their functions through metabolic tracing, immunofluorescence and protein–protein interaction assays. Lastly, we leveraged the atlas to generate a genome-scale screen of A-to-I RNA-editing modulators assayed through direct transcriptome-wide RNA editing, uncovering and mechanistically validating DBR1 as a potent regulator. Taken together, our data provide a comprehensive resource for interrogating the regulatory networks governing pluripotency, which is accessible at https://y-doctor.github.io/KOLF2.1J_Perturbation_Cell_Atlas/.

Mapping genotype–phenotype relationships is a fundamental goal of cell biology, providing critical insights into how a cell's genetic elements give rise to its cellular function. A cell type of particular interest is human induced pluripotent stem cells (hiPS cells), which possess the remarkable ability to differentiate into all somatic lineages, making them invaluable for developmental modeling, disease research and regenerative medicine applications^{1,2}. Substantial progress has been made in identifying key transcription factors and signaling pathways that regulate pluripotency and enable direct reprogramming^{3,4}. However, our understanding of pluripotency is far from complete, as the dynamic processes and gene interactions governing cell state transitions in stem cells are still not fully understood⁵. Systematic, large-scale functional genomic studies in hiPS cells have remained limited⁶ compared to those in cancer⁷, leaving key aspects of pluripotency regulation

less comprehensively mapped. We conservatively estimate that the functional roles of <2,000 genes have been deeply characterized within the stem cell context⁸ and hypothesize that additional, yet uncharacterized pathways may contribute to the regulation of pluripotency.

Therefore, we consider an expressed genome-wide perturbation cell atlas in hiPS cells as an invaluable resource toward understanding the pluripotent state. We generated a comprehensive CRISPR interference (CRISPRi) Perturb-Seq^{9–12} dataset spanning the entire expressed genome of KOLF2.1J hiPS cells¹³, accompanied by a parallel fitness readout. Through a custom analytical pipeline, we were able to resolve a map of functionally related genes, showing significant correlative enrichment of known regulatory modules and protein–protein interactions (PPIs). We validated key perturbations using orthogonal approaches, including size-exclusion chromatography coupled to mass

spectrometry (SEC–MS) and proximity-dependent biotin identification (BioID) for protein complex mapping, isotopolog metabolic tracing and RNA sequencing (RNA-seq). These validation experiments confirmed that transcriptomic clusters derived from Perturb-Seq reflect both interacting and functionally related protein complexes in hiPS cells. Notably, our approach uncovered putative metabolic (*ZBTB41*) and pluripotency (*RNF7*) regulators, providing insights into the roles of metabolic homeostasis and pluripotency in maintaining stem cell identity. In addition, we performed a transcriptome-wide analysis of A-to-I RNA editing across every CRISPR-perturbed KOLF2.1J population, generating an unbiased genome-scale screen of RNA-editing modulators assayed through a direct transcriptome-wide RNA-editing readout. We identified *DBRI* as a potent regulator, with its perturbation leading to significant increases in global RNA editing and deviation from basal cell state. Taken together, this dataset and its comprehensive validation provide a resource for dissecting pluripotency regulation.

Pan-expressed genome CRISPRi Perturb-Seq in KOLF2.1J hiPS cells

To optimize our experimental and computational methodologies, we selected KOLF2.1J hiPS cells, a well-established reference hiPS cell line¹³, and conducted two pilot CRISPRi Perturb-Seq screens targeting 108 chromatin modifiers and 97 metabolic enzymes respectively in KRAB^{ZIM3}–dCas9 KOLF2.1J iPS cells (henceforth simply referred to as KOLF2.1Js) maintained in pluripotent maintenance medium (mTeSR)¹⁴ (Extended Data Fig. 1a and Supplementary Note 1). Each gene was targeted with six independent sgRNAs along with 109 non-targeting control (NTC) sgRNAs, which served as a proxy for the unperturbed state. For clarity, a ‘perturbation’ represents an individual sgRNA, whereas a ‘perturbed gene’ represents the aggregate profile over all sgRNAs targeting the same gene. To identify ‘strong perturbations’ (perturbed genes that significantly alter cell state relative to NTC), we developed a custom computational pipeline that aimed to balance resolution with computational runtime because of the massive scale of the data. Full details and the rationale behind parameter choices can be found in the Methods and Supplementary Figs. 1 and 2; a diagram of the computational pipeline is presented in Extended Data Fig. 2. In brief, after isolating single, live cells harboring a single perturbation, we further select for perturbations inducing at least 30% on-target knockdown. Next, we perform both energy distance (Edist) testing¹⁵ and differential expression (DE) analysis¹⁶ using NTCs from the same batch as the reference for each perturbation. Edist testing provides a statistical measure of the perturbation-induced global shift in expression profile. DE analysis provides a local view, identifying individual genes that change in expression upon perturbation (false discovery rate (FDR): 0.05; Methods). Individual perturbations targeting the same gene with significant results for both these metrics are then combined per gene and the Edist and DE tests are repeated per perturbed gene. Those perturbed genes that are significant in this second round of testing are identified as strong perturbations, having demonstrated statistically significant perturbations relative to the unperturbed state, and kept for downstream analyses.

On the chromatin modifiers dataset, 92% of perturbations had >25 associated single cells and 65% of perturbations achieved >30% on-target knockdown (Extended Data Fig. 1b,c). Our custom computational pipeline recapitulated biologically meaningful results in the pilot screens such as *DNMT1* and *UHRF1* as top perturbations, with significant transcriptional impacts measured by both Edist (Extended Data Fig. 1d,e) and DE (Extended Data Fig. 1f,g). Further, these perturbations demonstrated strong co-clustering (Extended Data Fig. 1h,i), visualized at both the aggregate (Extended Data Fig. 1j and the single-cell (Extended Data Fig. 1k) level, consistent with their known interaction in DNA methylation maintenance^{17,18} (Extended Data Fig. 1l). In the metabolic enzymes dataset (Extended Data Fig. 1m–q), *MTOR* was identified as a top

perturbation (Extended Data Fig. 1o), consistent with its known involvement in metabolic regulation¹⁹.

We next scaled our efforts to generate a comprehensive Perturb-Seq dataset of KOLF2.1J hiPS cells with an expressed genome-wide CRISPRi library targeting 11,692 unique genes (transcripts per million mapped reads (TPM) > 1.5) (Fig. 1a). For each gene, three sgRNAs were designed by CRISPick^{20,21}, along with a total of 478 NTC sgRNAs. KOLF2.1J cells maintained in mTeSR were harvested on day 6 after transduction and processed for single-cell (sc)RNA-seq. A total of 5.3 million unique cell barcodes were detected (Supplementary Table 1). After filtering for live cells harboring a single perturbation, our approach yielded more than 2.5 million single cells with a median of >5,000 unique molecular identifiers (UMIs) per cell. Using the same cell population, we simultaneously carried out a CRISPR fitness screen²², harvesting bulk genomic DNA on day 6 and further on day 14 to assess the impact of gene knockdown on cell fitness (Fig. 1b).

The fitness screen revealed that, on day 6, 1,162 genes were significantly depleted (fitness z score < –1), increasing to 1,645 by day 14, indicating the progressive identification of essential genes. Similarly, the number of enriched genes (fitness z score > 1) increased from 77 on day 6 to 205 on day 14 (Fig. 1c and Supplementary Table 2). A functional enrichment analysis (Extended Data Fig. 3a) connected enriched perturbations to chromatin remodeling (SAGA complex components) and pathways related to apoptosis and DNA damage response (Fig. 1d, left). This suggests that many of these genes may function as tumor suppressors, where their disruption could confer a fitness advantage. For instance, *TP53*, *BAX*, *BAK1*, *PMAIP1* and *BBC3* (also known as *PUMA*) are well-established apoptosis and DNA damage response genes that, when depleted, are known to enhance proliferation and survival in pluripotent stem cells^{23–26}. The SAGA complex, in particular, has also been previously identified as an important regulator in pluripotent stem cells²⁷. This highlights the critical role of chromatin remodeling and apoptosis pathways in modulating cellular fitness within iPS cells. To explore functional enrichment of depleted perturbations (referring to essential genes), we overlapped the essential gene list in KOLF2.1Js with the essential gene list from the DepMap database²⁸, which was generated from CRISPR screens across 1,150 cancer cell lines (release 24Q2), revealing 1,202 shared genes predominantly encoding core cellular machinery such as cell cycle and translation (Fig. 1d, left, and Extended Data Fig. 3d,e). The essential gene list exclusive to KOLF2.1Js was enriched for pathways related to mitochondrial translation and Fanconi anemia pathway, which underscores the sensitivity of iPS cells to metabolic stress and DNA damage repair mechanisms. Using STRING PPI tools, we analyzed PPI networks of top depleted and enriched perturbations, revealing functional interconnectivity (Fig. 1d, right) and shared fitness impacts (Fig. 1e). Perturbations to members of the mitochondrial translation and Fanconi anemia clusters were progressively depleted from day 6 to day 14, indicating essential roles for iPS cell survival, while perturbations to apoptosis regulation and SAGA complex cluster members were enriched, suggesting fitness advantages. Individual gene fitness z scores within these clusters are detailed in Extended Data Fig. 3f,g. When comparing results from our screen to those in a recent hiPS cell CRISPRi fitness study by Usluer et al.²⁹, a strong correlation ($r = 0.713$, $P < 0.0001$) was observed (Extended Data Fig. 3l), along with shared essential pathways (Extended Data Fig. 3m,n), supporting the robustness and reproducibility of the screen.

We then analyzed the genome-scale Perturb-Seq dataset using the computational pipeline we optimized on the small-scale screens to further investigate the phenotypic effect of each perturbation. On average, 88% of perturbations had a coverage of over 25 associated cells (Fig. 1f), and 80% achieved greater than 30% mean target knockdown efficiency (Fig. 1g). We validated 24 perturbations through single-plex qPCR ($P < 0.0001$; Extended Data Fig. 3h–k), corroborating the perturbation efficacy observed from the single-cell experiment. Further,

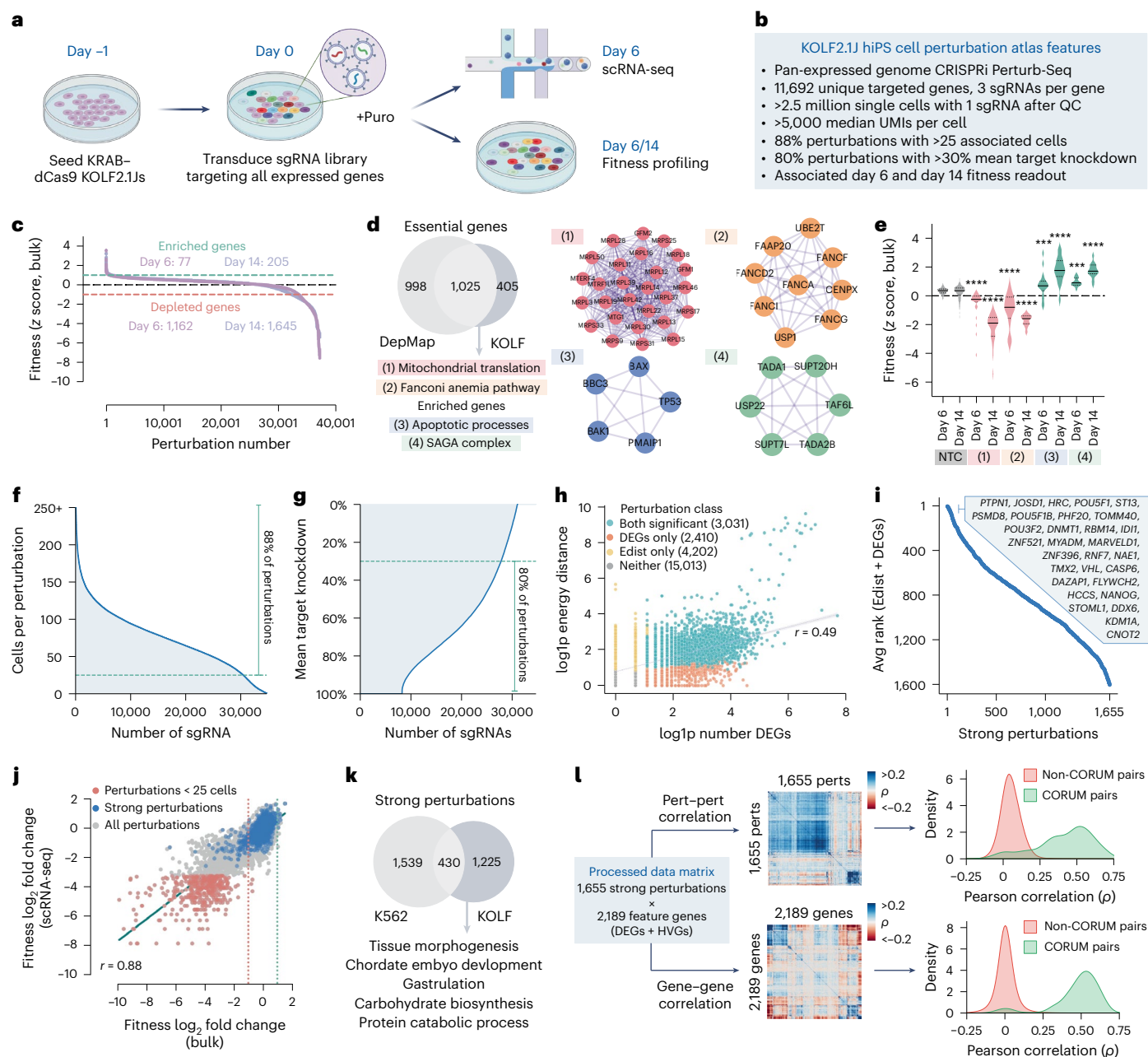


Fig. 1 | KOLF2.1J perturbation cell atlas. **a**, Experimental scheme. KRAB^{ZIM3}-dCas9 KOLF2.1J hiPS cells were seeded on day -1, transduced with an sgRNA library targeting all expressed genes on day 0, selected with puromycin starting on day 2 and harvested for single-cell Perturb-Seq on day 6. In parallel, cells were harvested on day 6 and day 14 for fitness profiling. Created with BioRender; Nourredine, S. <https://biorender.com/vsa8wre> (2026). **b**, Dataset overview. This genome-scale CRISPRi Perturb-Seq dataset covers 11,692 unique targeted genes with three sgRNAs per gene, along with 478 NTC sgRNAs across >2.5 million single cells. **c**, Waterfall plots of genes ranked by fitness z score in descending order on days 6 and 14. Enriched and depleted genes are identified by $z > 1$, $P < 0.05$ using MAGeCK⁹². **d**, Top, Venn diagram representing overlap between depleted perturbations in KOLF2.1J and the common essential genes identified in 1,150 cell lines from DepMap²⁸ portal (release 24Q2). Bottom, list of pathways exclusively depleted and enriched in KOLF2.1J. Right, PPI enrichment analysis of the top depleted and enriched perturbations. **e**, Distribution of fitness z scores on day 6 and day 14 of 478 unique NTC sgRNAs or those related to (1) mitochondrial translation (pink; 30 genes, 90 sgRNAs); (2) Fanconi anemia pathway (orange; 12 genes, 36 sgRNAs); (3) positive regulation of apoptosis (blue; 7 genes, 21 sgRNAs); or (4) SAGA complex (green; 6 genes, 18 sgRNAs). **f**, Number of cells per perturbation (sgRNA) captured by scRNA-seq, indicating that 88% of perturbations have >25 associated cells. **g**, Distribution of mean

on-target knockdown efficiency, indicating that 80% of perturbations achieve >30% knockdown. **h**, Relationship between Edist and number of induced DEGs per perturbation. Perturbations with significance for both metrics are retained. sgRNAs are combined per gene to form strong perturbations, defined as genes passing significance thresholds (Methods). **i**, Waterfall plot ranking strong perturbations on their average combined rank of Edist and DEGs, with the top 30 perturbations labeled. **j**, Correlation between fitness calculated from bulk genomic DNA extraction on day 6 versus fitness calculated from abundance of cells captured by scRNA-seq. Each point represents a perturbation; points in gray represent all perturbations detected with >1 cell (11,688), points in orange represent perturbations detected with <25 cells (427) and points in blue represent strong perturbations (1,665). **k**, Venn diagram representing overlap between strong perturbations in K562 cells (Replogle et al.³⁶) and KOLF2.1J, with labeling pathways enriched exclusively in KOLF2.1J. **l**, A set of 2,189 feature genes was chosen to represent the transcriptional profiles of strong perturbations (Methods). Top, pairwise correlation matrix of pseudobulked transcriptional profiles of strong perturbations and KDE plot of correlation distribution between pairs of strong perturbations within the same CORUM complex (green) and non-CORUM pairs (red). Bottom, pairwise correlation matrix of feature genes and KDE plot of correlation distribution between pairs of feature genes within the same CORUM complex (green) and non-CORUM pairs (red).

for 15 pluripotency markers, we detected that the marker gene itself was also the gene decreased most in expression compared to NTC, demonstrating the efficacy and specificity of the engineered CRISPRi system (Supplementary Fig. 6a). A total of 3,031 perturbations were significant in Edist and DE testing, which were then combined per gene and further filtered down after a second round of Edist/DE testing to a set of 1,665 strong perturbations (Fig. 1h). The strong perturbations identified in the screen highlight well-established pluripotency or developmental factors, such as *POU5F1* (also known as *OCT4*)³, *NANOG*³⁰, *PTPN1* (also known as *PTP1B*)³¹, *ZNF521* (also known as *EZHf* or *evi3*)³², *PSMD8* (ref. 33) and *PHF20* (ref. 34), confirming the screen's capability in capturing factors with critical roles in maintaining the pluripotent state of hiPS cells (Fig. 1i and Supplementary Fig. 3b,c). We also note that the screen, as is perhaps expected, is unable to capture the effect of perturbations with strong negative fitness impacts (Fig. 1j) as there are insufficient cells captured for downstream analyses. The strong perturbations we identify are generally uncoupled from cell fitness (absolute fitness $\log_2$ -transformed fold change < 1). To further understand the functional identity of the strong perturbations, we performed pathway analysis³⁵ (Supplementary Fig. 3d) highlighting enrichment of chromatin-modification complexes and embryo development regulators. To broadly gain insights into cell-type-specific regulatory networks, we compared the KOLF2.1J strong perturbations to those observed in a K562 genome-scale Perturb-Seq experiment³⁶ by Replogle et al.³⁵ (Fig. 1k and Supplementary Fig. 3e–g). We identified 1,225 strong perturbations unique to KOLF2.1Js, which were enriched in pathways such as tissue morphogenesis, embryo development and gastrulation, suggesting cell-type-specific differences (more detailed comparison in the Methods).

To assess the phenotypic impact of the strong perturbations, we turned our attention to the full perturbation-by-cell-by-gene tensor (that is, three-dimensional matrix) of the dataset. A total of 37,567 genes (including intronic reads) were initially detected, from which we selected a set of 2,109 feature genes using a combination of highly variable genes (HVGs) and differentially expressed genes (DEGs) to establish a transcriptional signature (Methods). Strong perturbations were then pseudobulked per gene (that is, averaged over all cells) to collapse the tensor into a two-dimensional data matrix of 1,665 strong perturbations by 2,109 feature genes (Fig. 1l, left, and Supplementary Tables 3 and 4). We computed the pairwise Pearson correlation matrix for both strong perturbations and feature genes and noted a significantly higher correlation between pairs of strong perturbations/feature genes in the same CORUM complex versus non-CORUM pairs ($P < 0.0001$, Mann–Whitney U -test) (Fig. 1l, right). This highlights the capacity of the screen to resolve functionally interrelated genes in KOLF2.1Js—both genes whose perturbation induces similar transcriptomes (correlated strong perturbations) and genes that respond similarly across a diversity of perturbations (correlated feature genes).

Comprehensive mapping of core functional modules in KOLF2.1J hiPS cells

Leveraging the breadth of our genome-scale Perturb-Seq dataset, we performed unsupervised clustering of perturbation profiles to comprehensively map functional gene networks. To assess these relationships within our KOLF2.1Js, we took the strong perturbation correlation matrix and performed hierarchical density based clustering (HDBSCAN) to identify clusters of related genes (Fig. 2a). Using CORUM (a database of experimentally verified PPIs), STRING PPI (a database of both experimentally verified and predicted PPIs) and Gene Ontology (GO) functional enrichment analyses, we annotated the HDBSCAN clusters on the basis of the related functional role of contained strong perturbations. In tandem, to visualize the 1,665 dimensional correlation matrix in an interpretable format, we generated a two-dimensional (2D) minimum distortion embedding (MDE) of the matrix. Using Leiden clustering, we identified clusters in the 2D MDE attempting to maximize

congruence with the HDBSCAN clusters identified in 1,655 dimensions (Methods). The MDE can, therefore, be seen as an approximation of the correlation relationships in the high-dimensional space. It is important to note that Leiden clustering is a visualization-level clustering, which we use to illustrate relationships found within the strong perturbation correlation matrix. MDE and HDBSCAN clusters were then annotated using CORUM, STRING and GO functional enrichment analyses. In result, the annotated MDE corresponds to an expressed genome-wide functional genomics cell map of KOLF2.1J hiPS cells.

Among the strong perturbation HDBSCAN clusters were clusters related to signal transduction (for example, Hippo–YAP (C6), heparin sulfate proteoglycans (C7) and TGF β signaling (C18)), metabolism (for example, mitochondrial translation (C26) and transcription (C27)), epigenetic modification (DNA methylation (C12 and C15)) and cell division (Cohesin complex (C3)), among others (Fig. 2b, Extended Data Fig. 4b–d and Supplementary Tables 5–7). We observed two clusters related to pluripotency (C20 and C21), which contained the key pluripotency factors *POU5F1* and *NANOG*, respectively. Using the strong perturbations found within these two clusters, we created a DEG regulatory network (Fig. 2c) where each gene is a node and an arrow is drawn to all other genes that are DEGs upon knockdown of that gene. We observed that a large number of knockdowns of genes in these clusters downregulate *POU5F1* and upregulate *ZNF521* expression, reinforcing their critical role in stem cell maintenance^{32,37,38}. Further analysis of perturbations driving changes to marker genes for endoderm, ectoderm and mesoderm dermal lineages (Extended Data Fig. 4e) identified that perturbations from within these clusters drove similar transcriptional phenotypes, with a decrease in pluripotency and increase in endodermal marker genes. HDBSCAN clusters also displayed strong enrichment in protein complexes; we highlight some examples from the STRING PPI network outputs for neddylation, DNA methylation and the SAGA complex clusters (Fig. 2d). We similarly performed HDBSCAN clustering on the transpose of the data matrix, identifying 46 feature gene clusters (Extended Data Fig. 5a,b), finding clusters with enrichment for genes involved in a diversity of functions, including cellular architecture (for example, cytoskeleton (C0) and cell division (C2)), cellular signaling (for example, GTPase/acetylcholine signaling (C13), NODAL signaling (C26) and growth factor signaling (C35)), cellular stress response (C19, C24 and C25) and structural morphogenesis (C43, C44, C45 and C46), among others. When summarizing relationships between strong perturbations and feature gene clusters, it was observed that gene perturbations within the pluripotency strong perturbation cluster yielded an overall increase in the expression of genes within the structural morphogenesis feature gene cluster (Extended Data Fig. 5c). Additionally, comparison to existing Perturb-seq atlases (Replogle et al.³⁶ K562 and Feng et al.³⁹ hiPS cells; Methods) enabled classification of strong perturbation clusters as either generally shared across cell types (for example, mitochondrial translation clusters) or KOLF specific (for example, pluripotency clusters).

We further visualized the relationship between strong perturbations in a 2D MDE embedding as mentioned above (Fig. 2e). A majority of HDBSCAN cluster memberships were also preserved in the MDE clusters (Methods and Extended Data Fig. 6a,b). Clusters in the MDE were significantly enriched for CORUM, KEGG, GO and Reactome terms relative to randomly permuted clusters of the same size (Fig. 2f and Extended Data Fig. 6c,d), substantiating the atlas's capacity to group perturbations with shared molecular functions or protein interactions. To simplify the representation, we reduced MDE clusters to a single node with the node size corresponding to the number of member genes (Extended Data Fig. 6e). When coloring each node by their median day 14 fitness zscore, adjacent and functionally related complex clusters (for example, mitochondrial translation C14, C25, C10 and C20 on the map's left side) tended to exhibit similar values (Fig. 2g and Extended Data Fig. 6f,g). Similar trends were also observed on the basis of Edist and the number of DEGs (Fig. 2h and Extended Data Fig. 6h,i). These patterns revealed that some perturbation communities drive

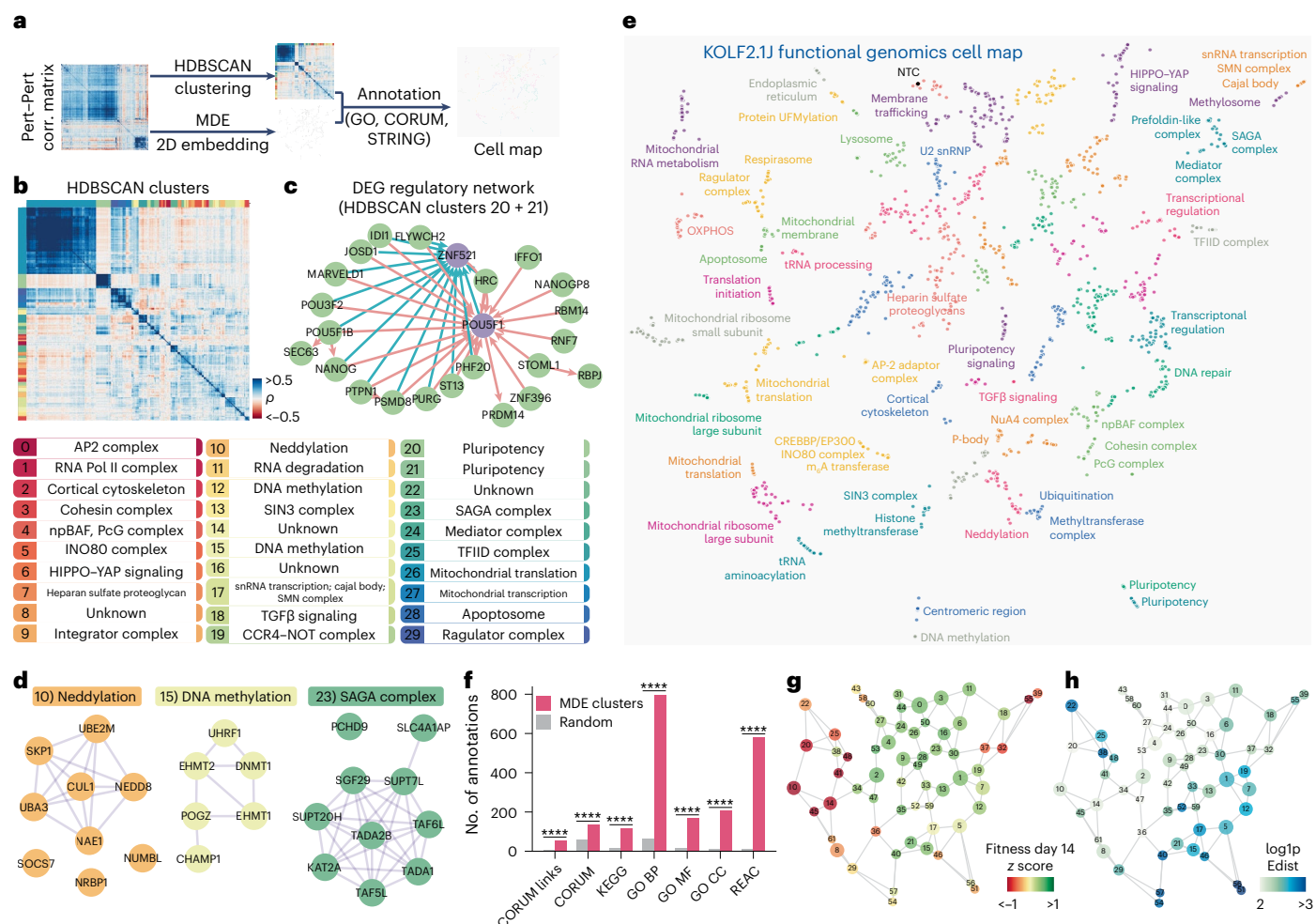


Fig. 2 | Systematic reconstruction of fundamental KOLF2.1J cellular processes.

a, Schematic of analysis methodology. Hierarchical clustering (HDBSCAN) is performed on the strong perturbation correlation matrix to identify and annotate clusters of perturbations with related functions. For visualization, the high-dimensional correlation matrix is embedded into two dimensions using an MDE. Clusters are then identified in the MDE using Leiden clustering with parameters chosen to retain HDBSCAN clusters (Methods). The 2D embedding is annotated using annotations from the HDBSCAN clusters, as well as known GO, CORUM and STRING terms. **b**, Top, strong perturbation correlation cluster map subset to only perturbations in HDBSCAN clusters. Bottom, annotation of HDBSCAN clusters based on functional enrichment analysis (Methods). **c**, Regulation network of pluripotency associated HDBSCAN clusters (20 + 21)

based on DE analysis. A salmon or teal arrow implies that the target node is a significant downregulated or upregulated gene, respectively, upon knockdown to the source node. **d**, STRING PPI networks of perturbations within selected HDBSCAN clusters. Edges are drawn for STRING scores > 0.9. **e**, Annotated MDE of strong perturbation correlation matrix, forming a functional genomics cellular map of KOLF2.1J hiPS cells. **f**, Functional enrichment analysis comparing the number of terms identified in MDE clusters to random clusters of the same size across annotation databases (**** $P < 0.0001$, two-sided Fisher's exact test). **g, h**, Each cluster of the MDE is collapsed to a single node, where the node size represents the number of cluster members. Edges are drawn between nodes and their three nearest neighbors. Nodes are colored by the median fitness day 14 z score (**g**) or median log1p Edist (**h**) of their associated members.

strong transcriptional remodeling (high Edist, large DEG numbers) without necessarily impacting cellular fitness, while others exert major effects on proliferation or survival with more modest transcriptional change. This decoupling between transcriptome remodeling and fitness highlights the value of combining transcriptional phenotyping with pooled fitness screening, enabling identification of distinct regulatory layers in cellular responses.

To experimentally assess whether transcriptional coclustering from Perturb-Seq reflects physical interactions, we performed SEC-MS in wild-type KOLF2.1Js. SEC-MS separates protein complexes by molecular weight and uses a guilt-by-association framework to assign coeluting proteins into the same complex^{40,41} (Extended Data Fig. 7a). The SEC-MS data were analyzed using the cosine similarity between fractionation profiles of proteins as physical evidence of their interactions (Methods). We first investigated strong perturbation HDBSCAN C29, containing LAMTOR complex members. C29 was predicted by

STRING to have strong PPIs (Extended Data Fig. 7b) and correspondingly showed strong cofractionation specifically in wild-type KOLF2.1Js by SEC-MS (Extended Data Fig. 7c,d). We similarly investigated strong perturbation HDBSCAN C26, containing mitochondrial translation machinery, as it contained the most STRING predicted PPIs of any HDBSCAN cluster we identified (Extended Data Fig. 7e). C26 contains members of five MDE clusters (C8, C10, C14, C20 and C45) and overlaying MDE cluster annotations onto the SEC-MS cosine similarity heat map revealed strong MDE intracluster and intercluster cofractionation patterns (Extended Data Fig. 7f), providing orthogonal validation of the Perturb-Seq-inferred C26 in KOLF2.1Js.

Identification of *ZBTB41* as a mitochondrial metabolism regulator

Perturb-Seq atlases are valuable tools for not only identifying regulatory complexes but also uncovering new gene functions. Our

analysis revealed distinct clusters of cofunctional strong perturbations, including one enriched in pluripotency factors (HDBSCAN C20) and another involving mitochondrial translation (HDBSCAN C26) (Fig. 3a). Within C26, we observed several known components of mitochondrial translation machinery including MRPL3, MRPL11, MRPL22, MRPL36, MRPL37 and MRPL58 (among others, as previously noted in Extended Data Fig. 4b). Furthermore, genes such as *ZBTB41*, *LFNG*, *KRT18*, *KPNA1*, *NAA40*, *SAMD10*, *C19orf84* and *PRR13* showed strong correlations with known mitochondrial factors, suggesting undiscovered roles in mitochondrial metabolism. We focused on *ZBTB41* as it is a poorly characterized zinc-finger transcription factor and, to validate its function, we transduced KOLF2.1J cells with sgRNA targeting *MRPL11*, *MRPL37* or *ZBTB41*. qPCR confirmed the knockdown of the genes (Fig. 3b). These knockdowns displayed fitness defects as was previously detected in the fitness screen (Fig. 1) and confirmed by cell growth assay (Fig. 3c). To explicitly trace the effect on knockdown on the tricarboxylic acid (TCA) cycle, cells were treated with [$^{13}\text{C}_6$] glucose or [$^{13}\text{C}_5$] glutamine for 24 h and the incorporation of labeled metabolites was measured using MS to probe metabolic consequences (Fig. 3d,e). Knockdowns of *MRPL11*, *MRPL37* and *ZBTB41* all resulted in a significant decrease in the succinate–fumarate ratio, along with alterations in the distribution of α -ketoglutarate (α -KG) isotopologs (Fig. 3f,g). This finding was further supported by a similar trend in isotopolog labeling for other key TCA cycle intermediates, including citrate, fumarate and malate (Supplementary Fig. 4).

Consistent with metabolic alterations, scRNA-seq, bulk RNA-seq and qPCR all showed upregulation of key mitochondrial transcripts MT-ND1, MT-ND2 and downregulation of MT-CO1 following each knockdown (Fig. 3h and Supplementary Table 8). Shared DEGs from bulk RNA-seq between the three knockdowns displayed a transcriptomic signature indicative of mitochondrial dysregulation (Fig. 3i). We further validated two additional *ZBTB41* sgRNAs (three total; Supplementary Fig. 5a–c). Comparative analysis with *MRPL11* and *MRPL37* identified 138 DEGs unique to *ZBTB41* (Supplementary Fig. 5d), enriched for oxidative phosphorylation and lipid response pathways (Supplementary Fig. 5e). Although *ZBTB41* is broadly expressed across cell lines (Supplementary Fig. 5f), DepMap²⁸ data indicate that it is selectively essential in only 12 of 1,186 cancer lines, suggesting a potentially distinct role in KOLF2.1Js (Supplementary Fig. 5g). Together, these functional assays, isotope-tracing experiments and transcriptome analyses pinpoint *ZBTB41* as a previously uncharacterized mitochondrial regulator whose loss impairs TCA cycle flux in KOLF2.1J hiPS cells.

Identification of *RNF7* as a pluripotency regulator

Interrogation of the pluripotency-enriched HDBSCAN C20 revealed several well-established pluripotency regulators including *POU5F1*, *PTPN1*, *ZNFS21* and *PHF20* (among others, as outlined in Extended Data Fig. 4b). We observed strong coclustering of ubiquitin ligase *RNF7* with these genes and noticed that *POU5F1* was downregulated upon *RNF7* knockdown from the Perturb-Seq data (Fig. 2c), suggesting it as a potential regulator of pluripotency. *RNF7* has not previously been directly linked to pluripotency regulation. We investigated the impact of *POU5F1* and *RNF7* perturbation across multiple assays (Fig. 3j–s, Extended Data Figs. 8 and 9i and Supplementary Figs. 6–8). qPCR confirmed efficient CRISPRi knockdown of *POU5F1* and *RNF7* versus sgNTC control (Fig. 3j) and bulk RNA-seq log₂-transformed fold changes for each perturbation were correlated ($r = 0.63$; Fig. 3k), with both transcripts significantly depleted (Fig. 3l). These findings were confirmed over two independent sgRNAs (Extended Data Fig. 8a–c), supporting an on-target effect. Consistently, western blotting confirmed efficient depletion of *POU5F1* and *RNF7* proteins upon knockdown (Fig. 3m). These results corresponded to morphological changes visible in culture (Fig. 3p and Extended Data Fig. 8d), where both sg*POU5F1* and sg*RNF7* cells adopted a flattened, differentiated-like morphology, indicative of pluripotency loss. To further assess their functional relationship,

we examined their expression at the single-cell level. In wild-type KOLF2.1J cells (that is, without an integrated CRISPRi system) (Fig. 3n), *POU5F1* protein localized to the nucleus while *RNF7* protein was nuclear or cytoplasmic. Quantification across more than 700 wild-type cells (Fig. 3o) showed a positive correlation ($r = 0.54$) between *RNF7* and *POU5F1* protein expression, suggesting that cells with high *RNF7* expression also retain higher levels of nuclear *POU5F1*. The sg*RNF7* KOLF2.1J cells (that is, with an integrated CRISPRi system) showed reduced *RNF7* signal accompanied by diminished nuclear *POU5F1*, except in some *RNF7*-positive cells (white arrows) that retained strong *POU5F1* expression (Fig. 3q). Consistent with these observations, single-cell quantification revealed that *POU5F1* protein levels decreased to near zero upon sg*POU5F1* knockdown and were significantly reduced upon *RNF7* depletion (Fig. 3r,s). By contrast, *POU5F1* depletion did not reduce *RNF7* expression, indicating that the regulatory relationship is likely unidirectional, with *RNF7* influencing *POU5F1* expression.

In addition, DE analysis identified substantial overlap in DEGs between both perturbations (615 shared DEGs; Extended Data Fig. 8e–g and Supplementary Table 8). To determine whether the regulatory interaction between *RNF7* and *POU5F1* extends to transcriptional programs indicative of pluripotency status, we analyzed bulk RNA-seq data from sg*POU5F1* and sg*RNF7* cells using gene set enrichment analysis (GSEA)^{42,43}. *POU5F1* knockdown produced a pronounced transcriptional shift, characterized by strong, significant enrichment for human ‘primitive endoderm cell embryo’ gene signatures (normalized enrichment score (NES) = 2.46) and concurrent depletion of ‘epiblast cell embryo’ programs (NES = –2.16). A similar, albeit milder, effect emerged following *RNF7* knockdown (primitive endoderm NES = 1.86; epiblast NES = –1.12), supporting the hypothesis that *RNF7* depletion promotes exit from pluripotency and biases toward endoderm lineage commitment, a pattern that has previously been observed following *POU5F1* loss^{44–47} (Extended Data Fig. 8h–j).

To further investigate potential mechanistic links between *POU5F1* and *RNF7*, we performed proximity-dependent biotinylation (BioID). We generated stable KOLF2.1J lines expressing GFP–BirA, *POU5F1*–BirA or *RNF7*–BirA fusion proteins under a doxycycline-inducible promoter, using the BioID2 system⁴⁸ for in situ biotinylation. Following 24 h of doxycycline induction, cells were lysed and biotinylated proteins were pulled down with streptavidin beads and identified by liquid chromatography–tandem mass spectrometry (LC–MS/MS) proteomics (Extended Data Fig. 9a,b). A total of 5,113 proteins (3,617 with ≥ 2 unique peptides) across the three conditions were identified (Extended Data Fig. 9c and Supplementary Table 9). Protein enrichment was calculated relative to the GFP–BirA control, which biotinylates proteins in both the nucleus and cytoplasm, providing a background reference. BioID of *RNF7* in KOLF2.1J revealed an enrichment of cytoplasmic and mitochondrial proteins (Extended Data Fig. 9f and Supplementary Fig. 6), including canonical ubiquitin ligase complex members such as CUL5 (refs. 49–51) and UBE2D2 (ref. 52) (Extended Data Fig. 9g). In contrast, the *POU5F1*–BirA interactome was enriched for nuclear proteins associated with transcriptional regulation (Extended Data Fig. 9j and Supplementary Fig. 7) and recovered known interacting complexes (Extended Data Fig. 9k) such as members of the MLL3–MLL4 complex⁵³ (Extended Data Fig. 9l), the β -catenin–TCF transcriptional module⁵⁴ (Extended Data Fig. 9m) and the SWI–SNF complex⁵⁵ (Extended Data Fig. 9n). We did not detect direct physical interaction between *RNF7* and *POU5F1* in the BioID datasets. The two proteins display distinct subcellular interactomes; *RNF7* predominantly engages cytoplasmic and mitochondrial ubiquitination machinery, while *POU5F1* associates with nuclear chromatin regulators (Supplementary Fig. 8). This suggests that the coregulation of pluripotency networks by *RNF7* and *POU5F1* is likely mediated indirectly (Supplementary Note 2). Together, these data provide a proximity-based interactome of *POU5F1* and *RNF7* in a human iPS cell model, offering a resource to explore their regulatory environment in pluripotent cells.

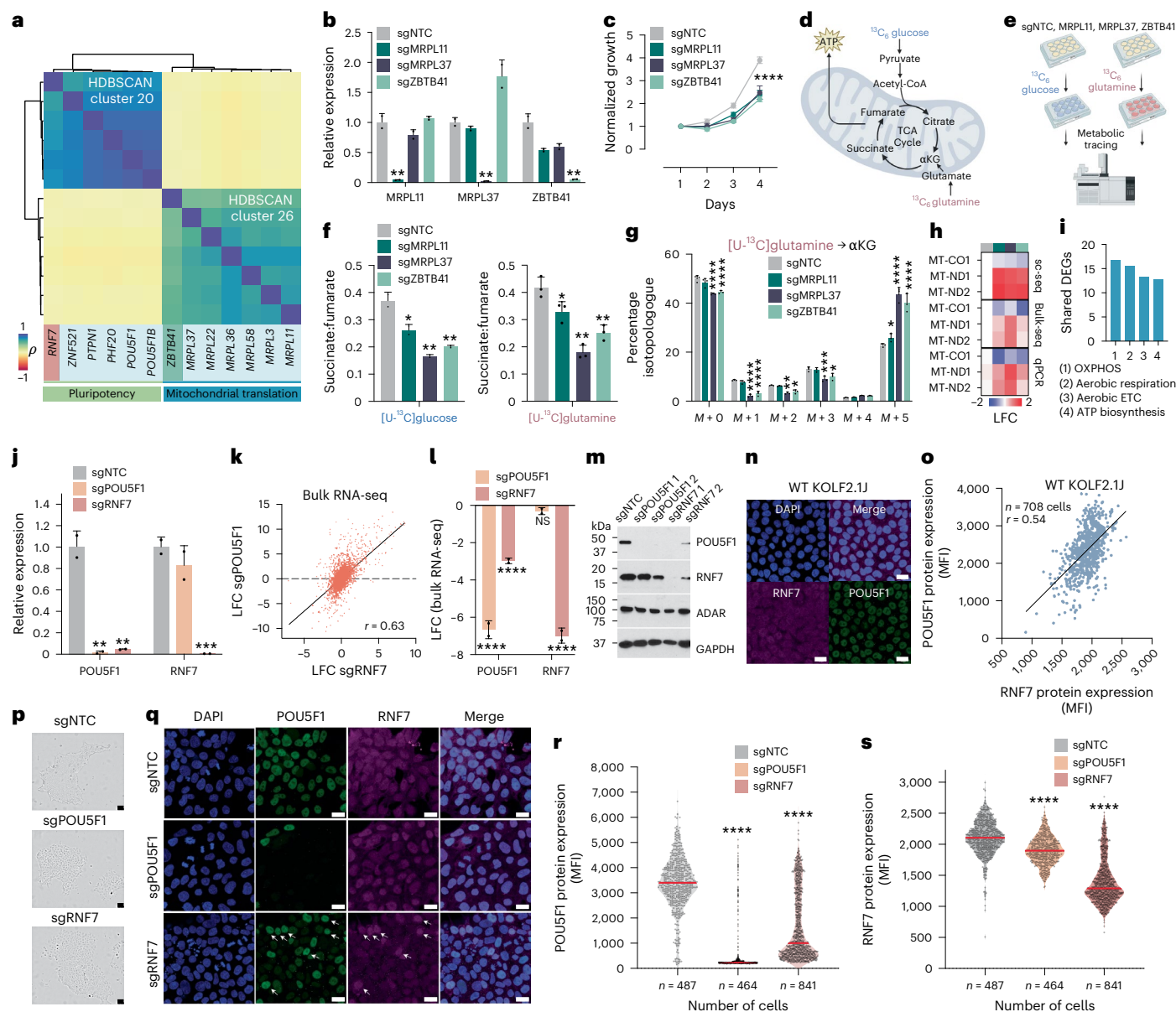


Fig. 3 | Discovery of metabolic factors and pluripotency regulators. **a**, Cluster map of selected strong perturbations from HDBSCAN C20 (pluripotency related) and C26 (mitochondrial translation related). Genes known from the literature to be involved in pluripotency and metabolism are highlighted in blue. RNF7 and ZBTB41 have not been previously characterized in pluripotency or metabolic processes. **b–i**, Validation experiments of putative metabolic factor ZBTB41. **b**, qPCR analysis of KRAB^{ZIM3}-dCas9 KOLF2.1J cells transduced with sgRNAs targeting NTC, MRPL11, MRPL37 and ZBTB41 ($n = 2$ biological replicates; $**P < 0.01$, two-tailed Student's t -test). **c**, CCK-8 cell growth assay, related to **b** ($n = 2$ biological replicates, $****P < 0.0001$, two-factor ANOVA; all conditions significant relative to NTC). **d, e**, Schematic of metabolic tracing array. ^{13}C -labeled glucose or glutamine was incubated with KRAB^{ZIM3}-dCas9 KOLF2.1J under each knockdown condition. For each member of the TCA cycle, isotopologues (M + 0 to M + 5) were measured by gas chromatography coupled to MS. Created using BioRender; Nourredine, S. <https://www.biorender.com/ergihlx> (2026). **f**, Ratio of succinate to fumarate traced by $[\text{U-}^{13}\text{C}]$ glucose (left) and $[\text{U-}^{13}\text{C}]$ glutamine (right) under knockdown conditions ($n = 2$ biological replicates; $*P < 0.05$ and $**P < 0.01$, two-tailed Student's t -test). **g**, Isotopologue analysis of the conversion of $[\text{U-}^{13}\text{C}]$ glutamine to α -ketoglutarate under knockdown conditions ($n = 3$ biological replicates; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$, two-factor ANOVA). **h**, Expression of a subset (MT-CO1, MT-ND1 and MT-ND2) of shared DEGs among MRPL11, MRPL37 and ZBTB41, assayed independently by scPerturb-Seq

(top), bulk RNA-seq (middle) and qPCR (bottom). **i**, GO terms associated with the union of all DEGs identified by bulk RNA-seq among MRPL11, MRPL37 and ZBTB41. **j**, qPCR analysis of KRAB^{ZIM3}-dCas9 KOLF2.1J cells transduced with sgRNAs targeting NTC, POU5F1 and RNF7 ($n = 2$ biological replicates). **k**, Correlation and Pearson correlation factor between bulk RNA-seq profiles of knockdown for POU5F1 and RNF7. **l**, Expression of POU5F1 and RNF7 from bulk RNA-seq profiles ($n = 2$ biological replicates). **m**, Western blot of KRAB^{ZIM3}-dCas9 KOLF2.1J knockdown for POU5F1 and RNF7. GAPDH was used as a loading control. Data are representative of $n = 3$ biological replicates with similar results. **n**, Immunofluorescence images of wild-type (WT) (no CRISPRi system) KOLF2.1J cells stained for RNF7 and POU5F1. Scale bars, 20 μm . **o**, Correlation and Pearson correlation factor between RNF7 and POU5F1 protein expression in wild-type KOLF2.1J cells quantified from $n = 8$ independent immunofluorescence images. MFI, median fluorescence intensity. **p**, Brightfield images representative of KRAB^{ZIM3}-dCas9 KOLF2.1J cells with knockdown of RNF7 and POU5F1. Scale bars, 75 μm . Enlarged images are present in Extended Data Fig. 8d. **q**, Immunofluorescence images of KRAB^{ZIM3}-dCas9 KOLF2.1J cells with knockdown of RNF7 and POU5F1 cells stained for RNF7 and POU5F1. Scale bars, 20 μm . **r**, Quantification of POU5F1 protein expression from $n = 9$ independent immunofluorescence images. **s**, Quantification of RNF7 protein expression from $n = 9$ independent immunofluorescence images ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$, two-factor ANOVA). Where error bars are present, data are presented as the mean values $\pm$ s.d.

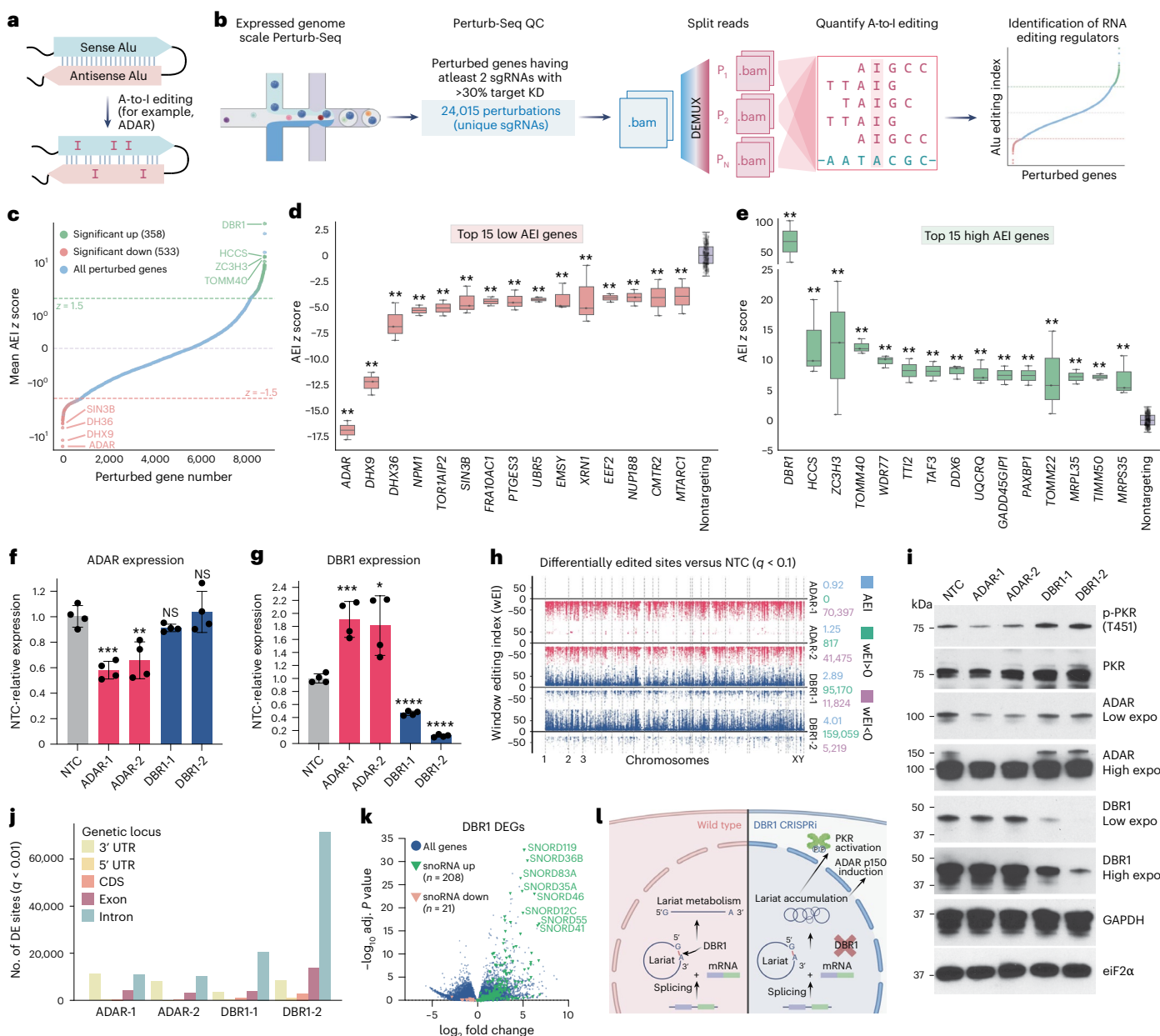


Fig. 4 | Unbiased genome-wide identification of RNA-editing regulators.

a, Schematic of Alu regions. Alu regions comprise ~10% of the human genome and are highly A-to-I edited regions of the epitranscriptome through ADAR enzyme. **b**, Analysis workflow. From the KOLF2.1J Perturb-Seq atlas, we identified 24,051 perturbations with >30% on-target knockdown and demultiplexed the pooled BAM file from scRNA-seq into individual files per perturbation. The AEI was then calculated for each perturbation. Created using BioRender; Nourredine, S. <https://biorender.com/g13tldz> (2026). **c**, Waterfall plot showing genes driving increased and decreased Alu editing relative to NTC. Significance was established through two-sided permutation testing versus NTC sgRNA adjusted using Benjamini–Hochberg correction (Methods; $^{**}P < 0.01$). **d, e**, Box-and-whisker plot showing the top 15 perturbed genes inducing low and high AEI from c. Each dot represents one sgRNA. Box plots show the median (center line), 25th–75th percentiles (box bounds) and whiskers extending to the minimum and maximum values. **f, g**, qPCR analysis of ADAR (f) and DBR1 (g) expression in KRAB^{ZIM3}–dCas9 KOLF2.1J cells transduced with sgRNAs targeting NTC, ADAR (two sgRNAs;

ADAR-1 and ADAR-2) and DBR1 (two sgRNAs, DBR1-1 and DBR1-2) ($n = 2$ biological replicates by $n = 2$ technical replicates; data are presented as the mean $\pm$ s.d.; NS, not significant; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$, two-tailed Student's *t*-test). **h**, Differential editing analysis by LoDEI⁶⁸ of A-to-I edits across the genome in cells transduced with sgRNAs targeting ADAR (pink) and DBR1 (blue). The AEI and number of differential sites with positive and negative window editing indices are also reported. **i**, Western blot analysis of KRAB^{ZIM3}–dCas9 KOLF2.1J transduced with sgRNAs targeting NTC, ADAR and DBR1. GAPDH was used as a loading control. Data are representative of $n = 3$ biological replicates with similar results. **j**, Number of differentially edited sites per condition in 3' untranslated region, 5' untranslated region, coding, exonic and intronic regions. **k**, DE analysis of bulk RNA-seq profile from DBR1 knockdown. snoRNAs are highlighted with a triangle. **l**, Proposed mechanism by which DBR1 knockdown leads to increased A-to-I RNA editing. Created using BioRender; Nourredine, S. <https://biorender.com/qvwrpvnv> (2026).

Unbiased genome-scale screen for regulators of A-to-I RNA editing

One of the underappreciated capacities of Perturb-Seq is that perturbation-induced changes are detectable both at the transcriptional

and post-transcriptional levels. While conventional analyses focus on changes in gene expression (Fig. 2), we sought to extend this capacity by systematically investigating whether perturbations also impact the nucleotide sequence of associated transcripts, specifically through

RNA-editing events. To explore this, we focused on A-to-I editing events, one of the most prevalent post-transcriptional modifications catalyzed by ADAR enzymes⁵⁶, which act broadly on double-stranded RNA (dsRNA) structures. Among endogenous sources of dsRNA, inverted Alu elements represent a major substrate for ADARs in human cells. These short interspersed nuclear elements form double-stranded RNA structures because of their inverted repeat architecture and comprise approximately 10% of the human genome⁵⁷, making them a dominant target class for A-to-I editing (Fig. 4a). This modification on dsRNA is critical in the modulation of innate immunity^{58,59}. To quantify transcriptome-wide A-to-I editing for each perturbation, we used the Alu editing index (AEI), which counts cumulative editing events occurring within these inverted Alu elements⁶⁰. Having previously demonstrated that AEI provides a sensitive and robust metric for measuring global RNA-editing activity⁶¹, even in low-read-coverage settings, we applied this approach to every perturbation in our genome-scale screen.

Using our Perturb-Seq atlas (filtered to perturbations passing basic quality control (QC) metrics), we systematically profiled the impact of ~24,000 perturbations on A-to-I RNA editing, using AEI as quantitative readout of global editing activity (Fig. 4b, Supplementary Table 10 and Methods). This results in a comprehensive expressed genome-wide screen of RNA-editing regulators with a transcriptome-wide readout⁶². We found that perturbations to 358 genes significantly increased A-to-I editing, while 533 significantly decreased editing (Fig. 4c). Notably, the top gene perturbation decreasing RNA editing was *ADAR*, consistent with its known function in facilitating A-to-I RNA editing (Fig. 4d). We also identified *DHX9* among the top perturbations reducing AEI. *DHX9* encodes a dsRNA helicase and is a functional interactor of *ADAR*^{63–65}. Conversely, for increased RNA editing, *DBR1* was the top gene by a factor of >5-fold relative to the next top hit (Fig. 4e). Several mitochondrial and RNA metabolism genes, including *TOMM40* and *HCCS*, also emerged as positive regulators. We note that a strong phenotypic effect of *DBR1* was also observed from the single-cell data, inducing over 1,000 DEGs (Supplementary Fig. 3b). Additionally from a DEG regulatory network of all strong perturbations, where genes were ranked by their regulatory effect using PageRank, *DBR1* was identified as a top two regulator (Extended Data Fig. 10a–c). While poorly characterized in the stem cell context, *DBR1* is known to encode an enzyme responsible for RNA lariat debranching and its knockdown has been shown to result in RNA lariat accumulation^{66,67}.

To validate our findings, we performed single-plex validation experiments on *ADAR/DBR1*, observing significant on-target knockdown (Fig. 4f,g). Total RNA-seq was then performed and differentially edited sites were identified by LoDEI⁶⁸ (version 1.0). We observed a genome-wide increase in A-to-I RNA editing upon *DBR1* knockdown and a decrease upon *ADAR* knockdown, with *DBR1* depletion generating tens of thousands of new edited sites predominantly localized in intronic regions (Fig. 4h,j and Extended Data Fig. 10d). Protein-level depletion was confirmed by western blot (Fig. 4i), showing loss of *DBR1* and *ADAR* upon knockdown. In *DBR1*-depleted conditions, we observed an upregulation of the *ADAR* p150 isoform and protein kinase R (PKR) phosphorylation (Thr451), both hallmarks of dsRNA sensing pathway activation^{69,70} (Fig. 4i). Furthermore, DEG analysis revealed upregulation of transcripts corresponding to small nucleolar RNAs (snoRNAs) in KOLF2.1J *DBR1*-depleted cells. This is consistent with prior findings showing that many box C/D and H/ACA snoRNAs are derived from intronic lariat precursors^{66,67}. In the iPS cell context, we found that, upon *DBR1* loss, cells exhibit transcriptional changes involving genes related to cell–cell adhesion, mitochondrial function and early developmental programs (Extended Data Fig. 10e). These patterns are consistent with alteration of cell identity, suggesting that loss of *DBR1* affects overall cell state. Taken together, these findings suggest that *DBR1* knockdown leads to RNA lariat accumulation, resulting in downstream *ADAR* p150 induction and PKR activation (Fig. 4i). Overall, our results identify *DBR1* as a previously unrecognized regulator of

RNA editing and innate immune signaling in KOLF2 hiPS cells, linking RNA lariat turnover to post-transcriptional control mechanisms and innate immune homeostasis.

Discussion

In this study, we performed an expressed genome-wide CRISPRi Perturb-Seq screen in KOLF2.1J hiPS cells, targeting 11,692 genes in parallel to systematically map the genetic architecture underlying iPS cell identity (Fig. 1). This Perturb-Seq atlas enabled the reconstruction of known molecular complexes and functional modules, while also revealing additional gene relationships that define the transcriptional and cellular landscape of KOLF2.1Js (Fig. 2). Beyond recapitulating established biology, the atlas proved effective for hypothesis generation and discovery, allowing us to identify previously uncharacterized regulators of cellular metabolism (*ZBTB41*) and pluripotency (*RNF7*), which we were able to experimentally validate (Fig. 3). Furthermore, by quantifying the effects of each expressed gene in KOLF2.1Js on A-to-I RNA editing across the transcriptome, we uncovered *DBR1* as a central modulator of dsRNA accumulation and innate immune sensing in pluripotent cells (Fig. 4), illustrating the broader potential of this resource to dissect both transcriptional and post-transcriptional gene function.

Through our clustering analysis, we identified distinct perturbation communities corresponding to known molecular complexes and regulatory modules, including chromatin regulators, mitochondrial translation machinery and key pluripotency factors. Functional clustering also revealed coclustering of uncharacterized genes. For instance, *ZBTB41* emerged from mitochondrial-associated clusters; metabolic flux analysis confirmed its role in mitochondrial metabolism, positioning *ZBTB41* as a regulator of TCA cycle activity. Regulation of the TCA cycle is particularly relevant in pluripotent stem cells, where metabolic state tightly interfaces with self-renewal, epigenetic remodeling and differentiation potential^{71–73}.

Our data also implicate *RNF7*, which encodes a RING-box protein and core component of Cullin–RING ligases (CRLs), as a regulator of pluripotency. While *RNF7* has been studied in cancer biology^{74,75}, its role in hiPS cells has not previously been defined. From our screen, *RNF7* coclustered with core pluripotency regulators and was experimentally validated as a modulator of *POU5F1* expression. While we did not detect direct physical interaction between *RNF7* and *POU5F1* in proximity-based proteomics, *RNF7* knockdown phenocopied *POU5F1* loss across multiple transcriptional and morphological assays, pointing to an indirect regulatory axis. CRLs are responsible for regulating a wide spectrum of cell-intrinsic processes, including cell-cycle progression⁷⁶ and DNA repair⁷⁷, many of which intersect with key features of the pluripotent state⁷⁸. Notably, *RNF7* also possesses antioxidant properties^{79,80}, comprising more than 10% cysteine, which facilitates thiol–disulfide exchange and metal chelation. *RNF7* has been shown to suppress copper-induced lipid peroxidation in vitro⁸¹. Given the sensitivity of pluripotent cells to oxidative stress⁸², it is plausible that *RNF7* contributes to stem cell maintenance also through redox buffering, potentially mitigating differentiation signals triggered by oxidative imbalance. Together, these results highlight that *RNF7* supports pluripotency networks in KOLF2.1J, illustrating how post-translational regulatory layers intersect with core transcriptional circuits to maintain stem cell identity.

Another major finding of this study is the identification of *DBR1* as a critical suppressor of innate immune activation and RNA-editing dysregulation in hiPS cells. Leveraging RNA capture at single-cell resolution, we quantified transcriptome-wide A-to-I RNA editing across all perturbations. *DBR1* knockdown resulted in a striking global increase in A-to-I editing, primarily localized to intronic regions, characterized by accumulation of lariat RNAs. These lariat RNAs form circular RNA structures and can adopt stable secondary structures that can be detected as dsRNA by cellular sensors^{83,84}. The resulting dsRNA accumulation leads to activation of the PKR, as evidenced by increased

phosphorylation of PKR at Thr451. We observed upregulation of ADAR p150, the interferon-inducible isoform of the A-to-I editing enzyme. This suggests that the cell, upon sensing excess dsRNA species generated by lariat buildup, compensates through ADAR p150 induction to buffer the immunostimulatory RNA load⁸⁵.

We acknowledge several limitations of this study, which we discuss in detail in Supplementary Note 3. In summary, because perturbations that strongly reduce fitness lead to rapid depletion of the affected cells over time, the Perturb-Seq assay is inherently biased toward capturing perturbations with minimal fitness costs. As a result, some essential regulators of pluripotency that impose strong negative fitness effects upon perturbation may be underrepresented. Furthermore, while the use of pluripotency maintenance medium (containing human fibroblast growth factor and transforming growth factor- β) was required to maintain an undifferentiated NTC population, this choice also required perturbations to overcome these strong self-renewal cues, potentially masking lineage-biasing effects. Although our sequencing depth (median ~5,000 UMIs per cell) and single-plex validations indicate robust CRISPRi repression, deeper profiling or use of dual-sgRNA designs could enhance signal strength and CRISPRi off-targets remain a potential confounder. Lastly, we strongly believe that continued development of more expressive and scalable analytic frameworks for genome-scale Perturb-seq datasets by the community will be critical to more comprehensively analyze these complex datasets.

Our dataset is a comprehensive resource for the scientific community to query the complexities of pluripotency, as demonstrated with several examples in this study. Future studies could include analysis of the dataset to focus on genes relevant for a differentiation trajectory of interest and investigating how perturbations push the undifferentiated stem cell toward that trajectory. For example, combining scRNA-Seq in a human-derived teratoma model⁸⁶ or a brain organoid⁸⁷ with perturbations would be an interesting method to assess and identify important genes involved in hiPS cell differentiation. Furthermore, we envision this dataset to be particularly useful for training models for both perturbation prediction⁸⁸ and hiPS cell state engineering⁸⁹, as a comprehensive Perturb-Seq dataset in a pluripotent stem cell line, providing data on the effect of perturbing every expressed gene. Lastly we anticipate great utility in combining datasets across perturbation cell atlases⁹⁰ to form a unified human cell perturbation atlas toward a comprehensive understanding of human cell biology⁹¹.

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/s41587-026-03199-w>.

References

- Yamanaka, S. Pluripotent stem cell-based cell therapy—promise and challenges. *Cell Stem Cell* **27**, 523–531 (2020).
- Park, I.-H. et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
- Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
- Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.* **22**, 410–424 (2021).
- Jia, Y. K., Yu, Y. & Guan, L. Advances in understanding the regulation of pluripotency fate transition in embryonic stem cells. *Front. Cell Dev. Biol.* **12**, 1494398 (2024).
- Li, K., Ouyang, M., Zhan, J. & Tian, R. CRISPR-based functional genomics screening in human-pluripotent-stem-cell-derived cell types. *Cell Genom.* **3**, 100300 (2023).
- Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C. & Vazquez, F. The present and future of the Cancer Dependency Map. *Nat. Rev. Cancer* **25**, 59–73 (2025).
- Joung, J. et al. A transcription factor atlas of directed differentiation. *Cell* **187**, 3161 (2024).
- Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
- 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).
- Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896 (2016).
- Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
- Pantazis, C. B. et al. A reference human induced pluripotent stem cell line for large-scale collaborative studies. *Cell Stem Cell* **29**, 1685–1702 (2022).
- Alerasool, N., Segal, D., Lee, H. & Taipale, M. An efficient KRAB domain for CRISPRi applications in human cells. *Nat. Methods* **17**, 1093–1096 (2020).
- Peidli, S. et al. scPerturb: harmonized single-cell perturbation data. *Nat. Methods* **21**, 531–540 (2024).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- Bostick, M. et al. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760–1764 (2007).
- Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**, 818–821 (2008).
- Saxton, R. A. & Sabatini, D. M. MTOR signaling in growth, metabolism, and disease. *Cell* **168**, 960–976 (2017).
- Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
- Sanson, K. R. et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.* **9**, 5416 (2018).
- Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
- Amir, H. et al. Spontaneous single-copy loss of TP53 in human embryonic stem cells markedly increases cell proliferation and survival. *Stem Cells* **35**, 872–885 (2017).
- Kim, E. J. Y. et al. BAK/BAX-mediated apoptosis is a Myc-induced roadblock to reprogramming. *Stem Cell Rep.* **10**, 331–338 (2018).
- Ilhry, R. J. et al. Genome-scale CRISPR screens identify human pluripotency-specific genes. *Cell Rep.* **27**, 616–630 (2019).
- Lake, B. B. et al. Context-dependent enhancement of induced pluripotent stem cell reprogramming by silencing Puma. *Stem Cells* **30**, 888–897 (2012).
- Naxerova, K. et al. Integrated loss- and gain-of-function screens define a core network governing human embryonic stem cell behavior. *Genes Dev.* **35**, 1527–1547 (2021).
- Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
- Usluer, S. et al. Optimized whole-genome CRISPR interference screens identify ARID1A-dependent growth regulators in human induced pluripotent stem cells. *Stem Cell Rep.* **18**, 1061–1074 (2023).
- Silva, J. et al. Nanog is the gateway to the pluripotent ground state. *Cell* **138**, 722–737 (2009).
- Matulka, K. et al. PTP1B is an effector of activin signaling and regulates neural specification of embryonic stem cells. *Cell Stem Cell* **13**, 706–719 (2013).

32. Kamiya, D. et al. Intrinsic transition of embryonic stem-cell differentiation into neural progenitors. *Nature* **470**, 503–509 (2011).
33. Szutorisz, H., Georgiou, A., Tora, L. & Dillon, N. The proteasome restricts permissive transcription at tissue-specific gene loci in embryonic stem cells. *Cell* **127**, 1375–1388 (2006).
34. Zhao, W. et al. Jmjd3 inhibits reprogramming by upregulating expression of INK4a/Arf and targeting PHF20 for ubiquitination. *Cell* **152**, 1037–1050 (2013).
35. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
36. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-Seq. *Cell* **185**, 2559–2575 (2022).
37. Shen, S., Pu, J., Lang, B. & McCaig, C. D. A zinc finger protein Zfp521 directs neural differentiation and beyond. *Stem Cell Res. Ther.* **2**, 20 (2011).
38. Kim, J. B. et al. Oct4-induced pluripotency in adult neural stem cells. *Cell* **136**, 411–419 (2009).
39. Feng, C. et al. A genome-scale single-cell CRISPRi map of trans gene regulation across human pluripotent stem cell lines. *Cell Genom.* **6**, 101076 (2026).
40. Fossati, A. et al. Toward comprehensive plasma proteomics by orthogonal protease digestion. *J. Proteome Res.* **20**, 4031–4040 (2021).
41. Fossati, A. et al. Next-generation proteomics for quantitative Jumbophage–bacteria interaction mapping. *Nat. Commun.* **14**, 5156 (2023).
42. Mootha, V. K. et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
43. 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).
44. Zafarana, G., Avery, S. R., Avery, K., Moore, H. D. & Andrews, P. W. Specific knockdown of OCT4 in human embryonic stem cells by inducible short hairpin RNA interference. *Stem Cells* **27**, 776–782 (2009).
45. Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.* **24**, 372–376 (2000).
46. Hay, D. C., Sutherland, L., Clark, J. & Burdon, T. Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells* **22**, 225–235 (2004).
47. Hough, S. R., Clements, I., Welch, P. J. & Wiederholt, K. A. Differentiation of mouse embryonic stem cells after RNA interference-mediated silencing of OCT4 and Nanog. *Stem Cells* **24**, 1467–1475 (2006).
48. Kim, D. I. et al. An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27**, 1188–1196 (2016).
49. Lumpkin, R. J., Baker, R. W., Leschziner, A. E. & Komives, E. A. Structure and dynamics of the ASB9 CUL–RING E3 Ligase. *Nat. Commun.* **11**, 2866 (2020).
50. Kostrhon, S. et al. CUL5–ARIH2 E3–E3 ubiquitin ligase structure reveals cullin-specific NEDD8 activation. *Nat. Chem. Biol.* **17**, 1075–1083 (2021).
51. Pang, E. Z. et al. Structural basis for L-isoaspartyl-containing protein recognition by the PCMTD1 cullin–RING E3 ubiquitin ligase. *J. Biol. Chem.* **301**, 110735 (2025).
52. Markson, G. et al. Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. *Genome Res.* **19**, 1905–1911 (2009).
53. Ding, J., Xu, H., Faiola, F., Ma'ayan, A. & Wang, J. Oct4 links multiple epigenetic pathways to the pluripotency network. *Cell Res.* **22**, 155–167 (2012).
54. Kelly, K. F. et al. β -catenin enhances Oct-4 activity and reinforces pluripotency through a TCF-independent mechanism. *Cell Stem Cell* **8**, 214–227 (2011).
55. Huang, X. et al. OCT4 cooperates with distinct ATP-dependent chromatin remodelers in naïve and primed pluripotent states in human. *Nat. Commun.* **12**, 5123 (2021).
56. Levanon, E. Y. et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22**, 1001–1005 (2004).
57. Levanon, K., Eisenberg, E., Rechavi, G. & Levanon, E. Y. Letter from the editor: Adenosine-to-inosine RNA editing in Alu repeats in the human genome. *EMBO Rep.* **6**, 831–835 (2005).
58. Zhang, Y., Bryant, J., Herron, L. & Mali, P. ADARs: pleiotropy in function, versatility in application. *Nucleic Acids Res.* **53**, gkaf672 (2025).
59. Rainaldi, J., Mali, P. & Nourredine, S. Emerging clinical applications of ADAR based RNA editing. *Stem Cells Transl. Med.* **14**, szaf016 (2025).
60. Roth, S. H., Levanon, E. Y. & Eisenberg, E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. *Nat. Methods* **16**, 1131–1138 (2019).
61. Dailamy, A. et al. Charting and probing the activity of ADARs in human development and cell-fate specification. *Nat. Commun.* **15**, 9818 (2024).
62. Wei, T. et al. Multimodal CRISPR screens uncover DDX39B as a global repressor of A-to-I RNA editing. *Cell Rep.* **44**, 116009 (2025).
63. Hong, H. et al. Bidirectional regulation of adenosine-to-inosine (A-to-I) RNA editing by DEAH box helicase 9 (DHX9) in cancer. *Nucleic Acids Res.* **46**, 7953–7969 (2018).
64. Cottrell, K. A., Andrews, R. J. & Bass, B. L. The competitive landscape of the dsRNA world. *Mol. Cell* **84**, 107–119 (2024).
65. Aktaş, T. et al. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115–119 (2017).
66. Buerer, L. et al. The debranching enzyme Dbr1 regulates lariat turnover and intron splicing. *Nat. Commun.* **15**, 4617 (2024).
67. Tang, Q., Liao, D. & Zheng, B. DBR1 orchestrates the fate of lariat RNA: debranching-dependent turnover and function. *Nucleic Acids Res.* **53**, gkaf639 (2025).
68. Torkler, P. et al. LoDEI: a robust and sensitive tool to detect transcriptome-wide differential A-to-I editing in RNA-seq data. *Nat. Commun.* **15**, 9121 (2024).
69. Jiao, H. et al. ADAR1 averts fatal type I interferon induction by ZBP1. *Nature* **607**, 776–783 (2022).
70. Lemaire, P. A., Anderson, E., Lary, J. & Cole, J. L. Mechanism of PKR activation by dsRNA. *J. Mol. Biol.* **381**, 351–360 (2008).
71. Ryall, J. G., Cliff, T., Dalton, S. & Sartorelli, V. Metabolic reprogramming of stem cell epigenetics. *Cell Stem Cell* **17**, 651–662 (2015).
72. Mathieu, J. & Ruohola-Baker, H. Metabolic remodeling during the loss and acquisition of pluripotency. *Development* **144**, 541–551 (2017).
73. Jackson, B. T. & Finley, L. W. S. Metabolic regulation of the hallmarks of stem cell biology. *Cell Stem Cell* **31**, 161–180 (2024).
74. Zhao, Y., Xiong, X. & Sun, Y. Cullin–RING ligase 5: functional characterization and its role in human cancers. *Semin. Cancer Biol.* **67**, 61–79 (2020).
75. Tao, Y. et al. Advancements and perspectives of RBX2 as a molecular hallmark in cancer. *Gene* **892**, 147864 (2024).
76. Jang, S.-M., Redon, C. E., Thakur, B. L., Bahta, M. K. & Aladjem, M. I. Regulation of cell cycle drivers by Cullin–RING ubiquitin ligases. *Exp. Mol. Med.* **52**, 1637–1651 (2020).

77. Hannah, J. & Zhou, P. Regulation of DNA damage response pathways by the Cullin–RING ubiquitin ligases. *DNA Repair (Amst.)* **8**, 536–543 (2009).
78. Wu, Y. & Zhang, W. The role of E3s in regulating pluripotency of embryonic stem cells and induced pluripotent stem cells. *Int. J. Mol. Sci.* **22**, 1168 (2021).
79. Guo, Y.-L., Chakraborty, S., Rajan, S. S., Wang, R. & Huang, F. Effects of oxidative stress on mouse embryonic stem cell proliferation, apoptosis, senescence, and self-renewal. *Stem. Cells Dev.* **19**, 1321–1331 (2010).
80. Sun, Y. & Li, H. Functional characterization of SAG/RBX2/ROC2/RNF7, an antioxidant protein and an E3 ubiquitin ligase. *Protein Cell* **4**, 103–116 (2013).
81. Swaroop, M. et al. Expression, purification, and biochemical characterization of SAG, a ring finger redox-sensitive protein. *Free Radic. Biol. Med.* **27**, 193–202 (1999).
82. Hu, Q. et al. Oxidative stress promotes exit from the stem cell state and spontaneous neuronal differentiation. *Oncotarget* **9**, 4223–4238 (2018).
83. Chen, Y. G. et al. Sensing self and foreign circular RNAs by intron identity. *Mol. Cell* **67**, 228–238 (2017).
84. Duan, C. et al. Defects in intron recycling suppress the antiviral response via a mechanism of intronic endogenous dsRNA. *J. Exp. Med.* **223**, e20250344 (2026).
85. Mannion, N. M. et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Rep.* **9**, 1482–1494 (2014).
86. McDonald, D. et al. Defining the teratoma as a model for multi-lineage human development. *Cell* **183**, 1402–1419 (2020).
87. Velasco, S. et al. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).
88. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935 (2024).
89. Magnusson, J. P. et al. PreciCE: precision engineering of cell fates via data-driven multi-gene control of transcriptional networks. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.11.04.621938> (2024).
90. Booesaghi, A. S., Galvez-Merchán, Á. & Pachter, L. Algorithms for a commons cell atlas. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.03.23.586413> (2024).
91. Rood, J. E., Hupalowska, A. & Regev, A. Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell* **187**, 4520–4545 (2024).
92. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).

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. 2026

¹Department of Bioengineering, University of California San Diego, San Diego, CA, USA. ²Department of Developmental Biology, Stanford University, Stanford, CA, USA. ³Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁴Molecular and Cell Biology Laboratory, The Salk Institute for Biological Studies, San Francisco, CA, USA. ⁵Quantitative Biosciences Institute (QBI), University of California San Francisco, San Francisco, CA, USA. ⁶Gladstone Institute of Data Science and Biotechnology, J. David Gladstone Institutes, San Francisco, CA, USA. ⁷Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, CA, USA. ⁸Division of Cellular and Clinical Proteomics, Department of Protein Science, SciLifeLab, KTH Royal Institute of Technology, Stockholm, Sweden. ⁹School of Medicine, University of California San Diego, San Diego, CA, USA. ¹⁰Big Data Institute, University of Oxford, Oxford, UK. ¹¹Department of Computer Science, The University of Alabama at Birmingham, Birmingham, VA, USA. ¹²Department of Medicine, University of Virginia, Charlottesville, VA, USA. ¹³Department of Pathology, Stanford University, CA, USA. ¹⁴Chan Zuckerberg Biohub, San Francisco, San Francisco, CA, USA. ¹⁵Department of Bioengineering and Therapeutics Sciences, University of California San Francisco, San Francisco, CA, USA. ¹⁶These authors contributed equally: Sami Nourredine, Yesh Doctor. ✉ e-mail: pmali@ucsd.edu

Methods

Cell culture

KOLF2.1Js (The Jackson Laboratory, JIPSC001000) were maintained under feeder-free conditions in mTeSR plus medium (StemCell Technologies). Before passaging, tissue-culture plates were coated with growth-factor-reduced Matrigel (Corning) diluted in DMEM/F-12 medium (Thermo Fisher Scientific) and incubated for 30 min at 37 °C, 5% CO₂. Cells were dissociated and passaged using the dissociation reagent Versene or Accutase (Thermo Fisher Scientific). HEK293T (American Type Culture Collection, CRL-3216) cells were cultured in DMEM (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific A5256701) and maintained at 37 °C in a humidified incubator with 5% CO₂. Cells were passed every 2–3 days upon reaching ~80% confluence using 0.05% trypsin–EDTA (Thermo Fisher Scientific) for dissociation.

KOLF2.1J genome-scale Perturb-Seq experiment

We engineered a KOLF2.1J line that stably expresses KRAB^{ZIM3}–dCas9 (refs. 14,36) and designed a CRISPRi sgRNA library that spans all protein-coding genes expressed at >1.5 TPM (assayed through bulk RNA-seq) in KOLF2.1Js. In total, 11,692 unique genes were targeted. To process the screen with reasonable workload, we had to split the screen into three different batches. Batch ALPHA contains a library of all annotated DNA-binding and DNA-modifying enzymes that were determined by combining existing studies and web-scraping of UniProt. Batches BETA and GAMMA contain all other genes expressed at >1.5 TPM, split evenly between the two batches. The library contains three sgRNA per gene target (35,076 sgRNA total) and 478 NTC sgRNAs, cloned into the CROP-Seq vector¹² and delivered to KOLF2.1Js by lentiviral infection at low multiplicity of infection (<0.3) on day 0. sgRNA library was ordered from Genscript. Starting on day 2, cells receiving perturbations were enriched by puromycin selection until day 6, at which point cells were harvested for scRNA-seq and day 6 fitness screening. Remaining cells were allowed to continue propagating to day 14, where they were harvested for day 14 fitness screening. For scRNA-seq, 115,000 transduced day 6 cells per channel were loaded onto a Next GEM Chip M single cell (10x Genomics); for each batch ALPHA, BETA and GAMMA, 32 channels across two Next GEM Chip M single cells were loaded for a total of 96 channels and samples were processed according to the manufacturer (10x Genomics, User Guide, [CG000416](#)). Gene expression libraries and downstream NGS were prepared and run according to the manufacturer Chromium Next GEM single-cell 3' HT reagent kits v3.1 (10x Genomics). Reagents generated for the screen are available for reuse upon contacting the corresponding author.

RNA extraction and real-time qPCR

Total RNA was extracted from cells using the RNeasy mini kit (Qiagen) according to the manufacturer's protocol. For cDNA synthesis, 1,000 ng of total RNA was reverse-transcribed using the ProtoScript first-strand cDNA synthesis kit (New England BioLabs), following the manufacturer's instructions. Real-time PCR qPCR was performed using SYBR green supermix (Bio-Rad) on a Bio-Rad CFX96 real-time PCR detection system. Gene expression levels were quantified relative to endogenous controls, with each sample measured in technical duplicates. The expression of target genes was normalized to the housekeeping gene *GAPDH* and data were analyzed using Bio-Rad CFX Manager software. Primers were ordered from IDT and sequences can be found in Supplementary Table 15.

Bulk RNA-seq library preparation and analysis

For transcriptome profiling following CRISPRi-mediated repression of ADAR and DBR1, total RNA was extracted from KOLF2.1J dCas9–KRAB cells using the RNeasy Plus mini kit (Qiagen) according to the manufacturer's instructions. For rRNA-depleted total RNA-seq, 500 ng of total RNA was subjected to rRNA removal using the NEBNext rRNA

depletion kit v2 (human/mouse/rat; New England Biolabs, E7400), followed by library preparation with the NEBNext Ultra II directional RNA library prep kit for Illumina (New England Biolabs, E7760). For poly(A) + RNA-seq, 500 ng of input total RNA was processed using oligo(dT) magnetic bead-based mRNA enrichment from the same NEBNext Ultra II RNA library kit workflow. Libraries were quantified using a Qubit fluorometer and size profiles assessed using TapeStation D1000 ScreenTape. Final libraries were sequenced on an Illumina NovaSeq X platform PE100. FASTQ files were aligned using BWA on hg38 human (*Homo sapiens*)^{93,94}. Gene expression was measured using Featurecounts on the aligned BAM files⁹⁵. Then, DE of each CRISPRi perturbation was calculated against sgNTC using DESeq2 (ref. 16). For RNA editing, differentially edited sites were determined for sgADAR-1, sgADAR-2, sg-DBR1-1 and sg-DBR-2 versus sgNTC by LoDE⁶⁸ using default parameters and hg38 human genome annotation (2024-A) with GENCODE v44 annotations. RNAEditingIndexer⁶⁰ was used to compute the AEI with default parameters and the built-in hg38 human genome reference, Alu annotations and single-nucleotide polymorphisms (SNPs).

Immunostaining and microscopy

For immunofluorescence staining, KOLF2.1J hiPS cells were seeded in 96-well plates and fixed with 4% paraformaldehyde (Thermo Scientific Chemicals, 043368.9M) in PBS for 15 min at room temperature. Following fixation, cells were permeabilized using 0.1% Triton X-100 (Sigma, T8787) in PBS for 15 min. Permeabilized samples were washed twice with PBS. Primary antibodies were diluted in PBS containing 4% FBS and applied at the following concentrations: mouse anti-OCT4 (Cell Signaling Technology, 75463T; 1:200), rabbit anti-RNF7 (Abcam, ab181986; 1:250), mouse anti-MYC (Cell Signaling Technology, 2276T; 1:1,000). Cells were incubated with primary antibodies overnight at 4 °C. The following day, cells were washed four times for 10 min with PBS and incubated for 90 min at room temperature with secondary antibodies diluted in PBS + 4% FBS: goat anti-rabbit Alexa Fluor 488 (Invitrogen, A11034; 1:800), goat anti-mouse IgG1 Alexa Fluor 555 (Invitrogen, A21127; 1:800) or goat anti-mouse IgG1 Alexa Fluor 647 (Invitrogen, A21240; 1:800), when applicable. Nuclei were counterstained with DAPI (Invitrogen, 62248) at 1.25 µg ml⁻¹ for 10 min. Lastly, cells were washed three additional times with PBS and stored in PBS at 4 °C before imaging. Imaging was performed using a Leica STELLARIS 8 confocal microscope equipped with an HC PL APO ×63/1.20 water-immersion CS2 objective (Leica Microsystems). Z stacks covering whole cells in three dimensions (up to 25 planes spaced 0.7 µm apart) were acquired for each field of view. Maximum-intensity projection images were generated using Fiji (ImageJ).

For single-cell quantification, the AdipoQ ImageJ plugins⁹⁶ were applied to 25% scaled maximum-intensity projection images of each stack. Scaling was necessary to supply images that were compatible with the default StarDist nuclear segmentation model (called 'versatile (fluorescent nuclei)' and featured in the Fiji plugin of StarDist⁹⁷). StarDist settings were as follows: percentile normalization, 0–99.9%; probability/score threshold, 0.48; overlap threshold, 0.3; number of tiles, 1. AdipoQ Analyzer settings were as follows: increase range, false; minimum particle size, 10 µm² (125.04 pixels). AdipoQ results files were merged using a Python script to assemble a master table and further filtered to remove nuclear objects that were too small to represent a nucleus (minimum size of 300 pixels, representing ~24 µm²). Median fluorescence intensity refers to the median pixel intensity within the nuclear mask.

SEC–MS

SEC–MS was performed as previously described⁹⁸. In brief, wild-type KOLF2.1J cells were lysed in HNN buffer (50 mM HEPES pH 7 and 150 mM NaCl) supplemented with 0.5% NP-40 and protease and phosphatase inhibitor. After sonication to shred DNA, samples were cleared by

ultracentrifugation (40,000g for 15 min at 4 °C) and 1 mg of protein lysate (1 mg ml⁻¹) was loaded onto an Amicon column (30-kDa molecular weight cutoff; Millipore, UFC803096) to dilute NP-40 with three washes of HNN buffer, followed by concentration to a volume of 110 µl that was injected onto an HPLC system (Agilent, 1260 Infinity II LC system) connected to a BioSep SEC-s4000 size-exclusion column (Phenomenex). Each SEC run was performed using a flow of 0.5 ml min⁻¹ of HNN buffer and 5 fractions per min were collected 8.5 min after sample loading for a total runtime of 26 min. Fractions 7–78 (72 fractions in total) were further processed to be acquired by MS. Proteins in fractions were reduced by adding 100 µl of reduction buffer (8 M urea (4 M final), 50 mM HEPES pH 7 and 10 mM DTT (5 mM final)) and incubated at 37 °C for 45 min before alkylation by adding 10 mM final of iodoacetate and incubation for 30 min in the dark. Fractions were further diluted with 800 µl of 50 mM HEPES and incubated overnight after the addition of 1 µg of trypsin (Promega, V511X). The next day, samples were desalted with C18 96-well plates according to manufacturer instructions and elutions were vacuum-dried and stored at –80 °C until MS acquisition. Fractions were resuspended in 60 µl of 0.1% formic acid and 5 µl of 3 × 24 consecutive fractions were pooled to generate an ion mobility gas-phase fractionation library according to the method described previously⁹⁹. In brief, data were collected from 400 to 1,200 *m/z* using windows of 5 *m/z* with an overlap of 1 *m/z* on each side. The ion mobility range spanned 0.57–1.47 V cm⁻², with two quadrupole positions per ion mobility cycle. These windows were evenly distributed across seven acquisition methods, with each method comprising 15 ion mobility cycles. For the SEC fractions, the peptides were sprayed through a 20-mm ZDV emitter kept at 1,600 V and 200 °C. The MS instrument was operated in positive mode with a thermal ionization MS range 0.60 V s cm⁻² to 1.60 V s cm⁻². MS1 scans were collected from 100 to 1,700 *m/z*, while MS2 scans were collected using data-independent parallel accumulation–serial fragmentation (PASEF) acquisition with a range of 303–1,200 *m/z*, comprising 24 variable MS2 windows. Analysis of spectra were performed using Spectronaut (version 19.2.240905.62635) using default settings. Spectronaut protein-group-summarized intensities were used for all downstream SEC–MS analyses.

The full SEC–MS data matrix (8,251 proteins × 72 fractions) for each of the three samples was column-normalized using a robust local regression method to fit a smooth curve through the global intensity trends across each sample. The goal was to detect and preserve global protein intensity trends across fractions as a smooth curve and adjust protein intensities to this global trend to account for run-to-run noise in MS recovery efficiency. The first step in normalization was to calculate a summary intensity value per fraction as the column effect from a median polish procedure, using the R function *medpolish*, on the full log-transformed intensity matrix. Next, cubic polynomial curves were fitted to the summary data using exhaustive sliding windows 20 fractions wide with function *rlm* from the MASS package of R. The residuals of each fraction to all of the local regression curves that include the fraction were pooled and the median was taken as each fraction's normalization factor. All protein intensities in each fraction were then adjusted by the corresponding fraction's normalization factor by subtraction on log-transformed data. Protein intensity values were then transformed back to their non-log-transformed values and scaled per protein and per sample by dividing by the sum of each protein's intensity in a sample. To calculate a combined cosine similarity between protein elution profiles across the three replicates of 72 fractions each, the normalized matrices were concatenated fraction-wise into a global matrix of 8,251 proteins by 216 fractions. Missing values were set to 0. Cosine similarities were then calculated between all possible protein pairs as the similarity between vectors of length 216.

Enrichment of physical interactions detected by SEC–MS in Perturb-Seq-derived clusters was calculated by a comparison of the 95th percentile score of within-cluster cosine similarities to a random distribution of 95th percentiles calculated on 10,000 permuted

clusterings in the same data. Each permutation shuffled the gene labels in the cluster table such that cluster sizes remained the same per cluster name and the score of each cluster was compared to permutations on the same-named cluster. The portion of permutations with a 95th percentile score greater than or equal to the original cluster was used as the *P* value; *P* values were adjusted for multiple testing across the 51 clusters by the Benjamini–Hochberg FDR method and those clusters with FDR < 0.05 were considered significantly enriched.

Proximity labeling (BioID) and MS

For POU5F1 and RNF7, we cloned synthetic gBlocks encoding GFP–BirA, POU5F1–BirA and RNF7–BirA fusion constructs into a doxycycline-inducible PiggyBac expression vector containing a BioID2 and a MYC tag. Stable KOLF2.1J hiPS cell lines were generated by transposition and hygromycin selection. For proximity labeling, cells were treated with 1 µg ml⁻¹ doxycycline and 1 µM biotin for 24 h before harvest. Cells were lysed in ice-cold RIPA buffer supplemented with benzonase (250 U per sample) and protease inhibitors as described previously. Lysates were sonicated (three 10-s bursts) and clarified by centrifugation (13,000 rpm (15,871g), 20 min, 4 °C). Then, 10 mg of total protein lysate was incubated with 60 µl of prewashed streptavidin Sepharose beads for 3 h at 4 °C. Protein samples were diluted in TNE (50 mM Tris pH 8.0, 100 mM NaCl and 1 mM EDTA) buffer. RapiGest SF reagent (Waters) was added to the mix to a final concentration of 0.1% and samples were boiled for 5 min. TCEP was added to 1 mM (final concentration) and the samples were incubated at 37 °C for 30 min. Subsequently, the samples were carboxymethylated with 0.5 mg ml⁻¹ of iodoacetamide for 30 min at 37 °C followed by neutralization with 2 mM TCEP (final concentration). Proteins samples prepared as above were digested with trypsin (trypsin-to-protein ratio of 1:50) overnight at 37 °C. RapiGest was degraded and removed by treating the samples with 250 mM HCl at 37 °C for 1 h followed by centrifugation at 14,000 rpm (18,407g) for 30 min at 4 °C. The soluble fraction was then added to a new tube and the peptides were extracted and desalted using C18 desalting columns (Thermo Scientific, PI-87782). Peptides were quantified using BCA assay and a total of 1 µg of peptides were injected for LC–MS analysis. Trypsin-digested peptides were analyzed by ultrahigh-pressure LC (UPLC)–MS/MS using nanospray ionization. The nanospray ionization experiments were performed using a Tim-TOF 2 pro hybrid MS instrument (Bruker) interfaced with nanoscale reverse-phase UPLC (EVOSEP ONE). The Evosep method of 30 samples per day was applied using a reverse-phase column (10 cm × 150 µm) packed with 1.5-µm C18-beads (PepSep, Bruker) at 58 °C. The analytical columns were connected with a fused silica ID emitter (inner diameter: 10 µm; Bruker Daltonics) inside a nano electrospray ion source (Captive spray source, Bruker). The mobile phases comprised 0.1% formic acid as solution A and 0.1% formic acid + 99.9% acetonitrile as solution B. The MS settings for the TimSTOF Pro 2 were as follows: PASEF method for standard proteomics; mobility-dependent collision energy ramping, 95 eV at an inversed reduced mobility (1/*k₀*) of 1.6 V s cm⁻² and 23 eV at 0.73 V s cm⁻². Collision energies were linearly interpolated between these two 1/*k₀* values and kept constant above or below. No merging of TIMS scans was performed. Target intensity per individual PASEF precursor was set to 20,000. The scan range was set between 0.6 and 1.6 V s cm⁻² with a ramp time of 166 ms. A total of 14 PASEF MS/MS scans were triggered per cycle (2.57 s) with a maximum of seven precursors per mobilogram. Precursor ions in an *m/z* range between 100 and 1,700 with charge states of 3+ to 8+ were selected for fragmentation. Active exclusion was enabled for 0.4 min (mass width: 0.015 Th, 1/*k₀* width: 0.015 V s cm⁻²). Protein identification and label-free quantification were carried out using Peaks Studio X (Bioinformatics Solutions). For downstream analysis, only proteins with a –10 log₁₀*P* score corresponding to an FDR < 1% (as calculated by PEAKS) and supported by more than two unique peptides were retained. We used a spectral-count-based approach to quantify protein enrichment,

as described previously^{100–102}. The log₂ fold changes were calculated for each protein (>2 unique peptides) by comparing spectral counts in the POU5F1–BirA and RNF7–BirA samples to the GFP–BirA control, which served as a background reference for nonspecific biotinylation in both nuclear and cytoplasmic compartments.

Western blot. Cells were lysed in RIPA buffer supplemented with protease (Pierce Protease inhibitor, Thermo Fisher) and phosphatase inhibitors (Phostop, Roche). Protein concentrations were determined by BCA assay (Thermo Fisher). Equal amounts of protein were separated using Bio-Rad precast SDS–PAGE gels and transferred to PVDF membranes (Bio-Rad). Membranes were blocked in 5% BSA in TBST (TBS + 0.1% Tween-20) for 1 h at room temperature and then incubated overnight at 4 °C with primary antibodies diluted 1:1,000 in TBST + 5% BSA. The following primary antibodies were used: PKR (Cell Signaling Technology, 3072), phospho-PKR (T451) (Thermo Fisher, 44-668G), MYC tag (Cell Signaling Technology, 2276), streptavidin–horseradish peroxidase (HRP; Thermo Fisher, 21130), DBR1 (Thermo Fisher, PA5-30616), EIF2A (Cell Signaling Technology, 9722), ADAR (Cell Signaling Technology, 81284), GAPDH (Cell Signaling Technology, 2118), RNF7 (Abcam, EPRI2001) and POU5F1 (Cell Signaling Technology, 75463). After washing, membranes were incubated with rabbit or mouse HRP-conjugated secondary antibodies (Cell Signaling Technology, 7074 and 7076) for 1 h at room temperature, developed with enhanced chemiluminescence (34580) and imaged using an X-ray (Blue Devil) film and a manual film developer.

Metabolic tracing and mitochondrial flux

For isotopic labeling experiments, cells were cultured in mTESR containing 17.5 mM [U-¹³C₆]glucose or 2.5 mM [U-¹³C₅]glutamine (Cambridge Isotope Laboratories) for 24 h before extraction. To extract metabolites, cells were rinsed with saline solution. Then, 700 µl of ice-cold 60:40 methanol–water solution was added to the cells along with polar and fatty acid internal standards. Cells were scraped from the plate and lysates were transferred to Eppendorf tubes. Samples were vortexed for 5 min and 10% was taken for protein quantification by BCA. The remaining homogenate was combined with 500 µl of chloroform, vortexed for 5 min and centrifuged (5 min, 21,000g, 4 °C). The upper polar phase was collected, dried under a vacuum at 4 °C and derivatized with 2% (w/v) methoxyamine hydrochloride (Thermo Fisher Scientific) in pyridine for 60 min, followed by silylation in *N*-tert-butyldimethylsilyl-*N*-methyltrifluoroacetamide with 1% *tert*-butyldimethylchlorosilane.

Lentiviral overexpression of RNF7 in hiPS cells

Cells were maintained on Matrigel-coated plates in mTeSR plus medium (StemCell Technologies) under standard feeder-free conditions. RNF7 (UniProt [Q9UBF6](#)) was overexpressed by lentiviral transduction using a constitutive expression vector encoding human RNF7, with eGFP as a control. Transductions were performed at the same multiplicity of infection across all conditions. Then, 2 days after transduction, cells were selected with puromycin (0.75 µg ml⁻¹). RNA was isolated 2 days after puromycin selection. RNA-seq libraries were prepared using the NEBNext Ultra II RNA library prep kit (New England Biolabs).

Guide library generation

Targeted genes constituted all genes expressed at >1.5 reads per kilobase of million mapped reads from a bulk RNA-seq analysis of wild-type KOLF2.1Js. To target genes by CRISPRi, we selected the top three gRNAs and genes using CRISPRick. In total, we generated 35,217 gRNAs targeting genes and 478 nontargeting gRNAs. This pool of gRNAs was synthesized in three oligonucleotide pools (ALPHA, BETA and GAMMA; Genscript) and cloned into the CROP-Seq vector as described below.

The Gibson assembly reactions were set up as follows: 1,000 ng of digested backbone, 1:20 molar ratio of insert, 2× Gibson assembly

master mix (New England Biolabs) and H₂O up to 100 µl. After incubation at 50 °C for 1 h, the product was transformed by electroporation into One Shot Stbl4 Electromax chemically competent *Escherichia coli* (Invitrogen). A fraction (1 µl) of cultures was spread on carbenicillin (50 µg ml⁻¹) Luria–Bertani (LB) plates and incubated overnight at 37 °C to estimate coverage, while the remaining transformant was seeded into 150 ml of LB medium containing carbenicillin (50 µg ml⁻¹) and incubated overnight in a shaker at 37 °C for 16–18 h. The plasmid DNA was then extracted using a Plasmid maxi kit (Qiagen). sgRNAs used for validation experiments can be found in Supplementary Table 15 and were ordered as primers from IDT.

Library lentiviral production

HEK293T cells were maintained in high-glucose DMEM supplemented with 10% FBS. Cells were seeded in a 15-cm plate 1 day before transfection, such that they were 60–70% confluent at the time of transfection. For each plate, 36 µl of Lipofectamine 2000 (Thermo Fisher Scientific, 11668027) was added to 2 ml of Opti-MEM (Thermo Fisher Scientific, 31985062). Separately, 3 µg of pMD2.G (Addgene, 12259), 12 µg of pCMV delta R8.2 (Addgene, 12263) and 9 µg of the pooled gRNA vector library were added to 2 ml of Opti-MEM. After 5 min of incubation at room temperature, the Lipofectamine 2000 and DNA solutions were mixed and incubated at room temperature for 30 min. During the incubation period, the medium in each plate was replaced with 20 ml of fresh, prewarmed medium per well. After the incubation period, the mixture was added dropwise to each plate of HEK293T cells. Supernatant containing the viral particles was harvested after 48 and 72 h, filtered with 0.45-µm filters (Steriflip, Millipore) and further concentrated using Amicon Ultra-15 centrifugal ultrafilters (100,000 nominal molecular weight limit; Millipore) down to a final volume of 600–800 µl, for each 15-cm plate. Lastly, the concentrated supernatant was divided into 200-µl aliquots and frozen at –80 °C. Two 15-cm plates worth of lentivirus was adequate.

gRNA amplification

gRNA barcodes were amplified from cDNA generated by the 10x single-cell platform and gDNA from cells harvested on day 6 and 14 of the KOLF2.1J fitness screen. Amplicons were prepared for deep sequencing through a two-step PCR process.

For amplification of barcodes from cDNA, the first step was performed as four separate 50-µl reactions for each sample. Then, 2.5 µl of the cDNA was input per reaction with Kapa Hifi Hotstart ReadyMix (Kapa Biosystems, KK2602). The PCR primers used were as follows: CROPseq_sgRNA_barcodeAmp_F, GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTGTGGAAAGGACGAAACAC; NEBNext Universal_R, GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGTGGAAAGGACGAAACAC. The thermocycling parameters were 95 °C for 3 min, followed by 20–26 cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 30 s, with a final extension of 72 °C for 5 min. The numbers of cycles were tested to ensure that they fell within the linear phase of amplification. Amplicons (~450 bp) of four reactions for each sample were pooled, size-selected and purified with Agencourt AMPure XP beads (Beckman Coulter) at a 0.8 ratio. The second step of PCR was performed with two separate 50-µl reactions with 50 ng of first-step purified PCR product per reaction. NEBNext multiplex oligos for Illumina (dual index primers) were used to attach Illumina adaptors and indices to the samples. The thermocycling parameters were 95 °C for 3 min, followed by six cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 30 s, with a final extension at 72 °C for 5 min. The amplicons from these two reactions for each sample were pooled, size-selected and purified with Agencourt AMPure XP beads at an 0.8 ratio. The purified second-step PCR library was quantified by Qubit dsDNA HS assay (Thermo Fisher Scientific) and used for downstream sequencing on an Illumina NovaSeq platform.

For amplification of barcodes from genomic DNA, genomic DNA was extracted from stored cell pellets with a DNeasy blood and tissue

kit (Qiagen). The first-step PCR was performed as two separate 50- μ l reactions for each sample. Then, 500 ng of genomic DNA was input per reaction with Kapa HiFi Hotstart ReadyMix. The PCR primers used were as follows: sgRNAsequenceAmp_gDNA_F, AACTCTTTCCCTA-CACGACGCTCTCCGATCTTATATATCTTGTGAAGGACGAAACACCG; sgRNAsequenceAmp_gDNA_R, GACTGGAGTTTCAGCGTGTGCTCTTCGATCTCTTATTTAACTTGCTATTTCTAGCTCTA. The thermocycling parameters were 95 °C for 3 min, followed by 24–32 cycles of 98 °C for 20 s, 60 °C for 15 s and 72 °C for 30 s, with a final extension of 72 °C for 5 min. The numbers of cycles were tested to ensure that they fell within the linear phase of amplification. Amplicons (~150 bp) of the two reactions for each sample were pooled and size-selected with Agencourt AMPure XP beads at a ratio of 1.6. The second step of PCR was performed as two separate 50- μ l reactions with 25 ng of first-step purified PCR product per reaction. NEBNext multiplex oligos for Illumina (dual index primers) were used to attach Illumina adaptors and indices to the samples. The thermocycling parameters were 95 °C for 3 min, followed by 6–8 cycles of 98 °C for 20 s, 65 °C for 20 s and 72 °C for 30 s, with a final extension at 72 °C for 2 min. The amplicons from these two reactions for each sample were pooled and size-selected with Agencourt AMPure XP beads at a ratio of 1.6. The purified second-step PCR library was quantified by Qubit dsDNA HS assay (Thermo Fisher Scientific) and used for downstream sequencing on an Illumina NovaSeq platform.

Fitness screen analysis

CROP-Seq plasmid, day 6 and day 14 gRNA amplicon libraries (ALPHA, BETA and GAMMA) were sequenced on a NovaSeq X Sequencer (Illumina). FATSQ files were analyzed using the MAGeCK version 0.5.8 package for computing enrichment and depletion to the plasmid libraries at the gRNA and gene levels⁹². The fitness score corresponding to a z score was calculated as follows:

$$\text{Fitness score} = Z = (X - \mu) / \sigma$$

where X is the $\log_2$ -transformed fold change, μ is the mean of all $\log_2$ fold changes and σ is the s.d. of all $\log_2$ -transformed fold changes.

Statistics and reproducibility

All statistical analyses for validation experiments were performed using GraphPad Prism 9. For Fig. 3m,n,p, images shown are representative of three independent experiments. For Figs. 2f and 4 and Supplementary Fig. 4, statistical comparisons between conditions were conducted using a two-way analysis of variance (ANOVA) with multiple comparisons against the NTC. For qPCR experiments in Extended Data Fig. 3 and Fig. 4, comparisons were made using a two-tailed unpaired t -test against the NTC control condition. Statistical details, including exact P values and significance levels, are provided in the Source Data.

Perturb-Seq analysis pipeline

Genome reference alignment, cell calling, sgRNA assignment and aggregation. Genome reference alignment, cell calling, sgRNA assignment and channel aggregation were all performed using cellranger version 8.0.0 count and aggr using the cellranger human transcriptome reference (GRCh38)-2024-A and default command line arguments. Aggregation was performed on all channels with >5,000 median gene expression UMIs per cell and >15 median CRISPR UMIs per cell (corresponding to 90 of 96 channels) to ensure sufficient read depth and sgRNA call confidence for all channels with the --no-secondary flag. Aggregation was manually terminated after SC_RNA_AGGREGATOR. WRITE_MATRICES and SC_RNA_AGGREGATOR. CRISPR_ANALYZER. CALL_PROTOSPACERS chunks were completed.

Metadata assignment and cell-level quality control metrics. The analysis pipeline works directly on the output of cellranger aggr. First,

we read the filtered_feature_bc_matrix.mtx.gz data matrix generated by cellranger using `anndata.read_h5ad()` into an `Anndata` object. For each cell, the corresponding sgRNA is assigned through the `protospacer_calls_per_cell.csv` file produced as cellranger aggr output. Cells without exactly one sgRNA assigned are removed (Extended Data Fig. 2a and Supplementary Fig. 1a,b).

After isolating cells with exactly one sgRNA assigned, we assigned metadata per cell (run ID, 10x chip ID, 10x channel ID, batch ID (ALPHA, BETA or GAMMA), cell type (KOLF2.1J) and perturbation type (CRISPRi)). Genes with exactly zero UMIs detectable in the entire dataset were removed (Extended Data Fig. 2b). Dead and aberrant cells were removed using the median absolute deviation (MAD) method described previously¹⁰³, filtering on five MADs for the QC covariates: `log1p_total_counts`, `log1p_n_genes_by_counts`, `pct_counts_in_top_20_genes` and `pct_counts_mt` evaluated by `scanpy's sc.pp.calculate_qc_metrics()` with cumulative proportion to the 20th most expressed gene (`percent_top = [20]`) (Extended Data Fig. 2c,d).

Core control cell identification. Although NTC sgRNAs are designed to not directly target any genomic loci, there are instances where an NTC sgRNA, perchance, may induce a phenotypic alteration³⁶. In light of this, we sought to remove outlier NTC sgRNA and NTC cells that demonstrated phenotypic deviation from the norm, with the underlying assumption that NTC sgRNAs typically do not induce phenotypic alterations compared to wild-type KOLF2.1J hiPS cells. To do so, we leveraged the results of fitness screening, removing any NTC sgRNA that was not within the middle 50% interquartile range of fitness $\log_2$ fold changes. This corresponded to a NTC sgRNA 'whitelist' of 236 sgRNAs with mean fitness $\log_2$ fold change of 0.3, max $\log_2$ fold change of 1.0 and min $\log_2$ fold change of -0.4. NTC sgRNAs in the aggregated `anndata` object were subset down to contain only those in the whitelist. We then used an outlier detection model (isolation forest¹⁰⁴) to filter out NTC cells with a deviant phenotype from the norm (Extended Data Fig. 2e). NTC cells were isolated from the perturbed cells and batch-aware HVG selection was performed using `sc.pp.highly_variable_genes` with `flavor = 'seurat_v3'` to select 2,000 HVGs, embedded through the top 50 principal components, batch-aligned in PCA space using Harmony¹⁰⁵ and filtered using an isolation forest (sklearn) with a 30% contamination fraction. The choice of 30% contamination fraction and the middle 50% interquartile range are conservative values we selected, as the true proportion of deviant NTC cells is unknowable. A potential limitation of this two-stage NTC pruning approach is potential exclusion of genuine biological variability within the NTC population, although we confirmed that this filtering did not substantially inflate downstream test statistics relative to an approach without NTC filtering (Supplementary Fig. 1d). We anecdotally noted that, from a uniform manifold approximation and projection (UMAP) visualization of the NTC cells, choosing a contamination fraction of 30% removed a majority of cells 'projecting outward' from the core population (Extended Data Fig. 2e, middle). When applying this pipeline to other datasets, these parameters may need to be adjusted. Remaining NTC cells after this filtering were identified as core NTC cells and merged back with the perturbed cells.

sgRNA and cell-level quality control. After core NTC detection, counts per cell were normalized to 1 million. The following perturbations were filtered out: (1) sgRNAs that did not induce, on average over all associated cells, >30% on-target knockdown; (2) sgRNAs without at least 25 associated cells; and (3) single cells that did not experience any on-target knockdown. NTC sgRNAs without at least 25 remaining cells are also removed. The thresholds of 30% on-target knockdown and 25 cells were chosen as they were the parameters used by Replogle et al.³⁶. Target knockdown was evaluated relative to the mean target expression within NTC cells from the same batch. Furthermore, we removed any perturbed cells that were not repressed to any level for the sgRNA they received (that is, they did not have on-target knockdown > 0%) (Extended Data Fig. 2f).

Perturbation-level Edist and DE testing. To assess which perturbations induced significant alterations to cell state, we leveraged both Edist testing^{15,36} and DE analysis¹⁶. Briefly, the Edist measures the distance between two population distributions, accounting for the variability within each population. Two populations with large pairwise distances between their respective members and small pairwise distances within their members have a high Edist. By performing permutation testing on the Edist metric, a *P* value for significance can be assessed. An excellent analysis of how this metric can be applied to Perturb-Seq data is presented by Peidli et al.¹⁵. An orthogonal metric to assess perturbation strength is to compute the number of DEGs it induces. While the Edist assesses global shifts in transcriptome, DEGs assess local shifts (that is, at the per-gene level). We sought to filter for perturbations that induce both significant global and local shifts to the cell's transcriptome.

We performed Edist testing between each perturbed cell population and NTC cells from the same batch. While Peidli et al.¹⁵ recommend generally having at least 200 cells per population, they also acknowledge that, in 'real-world datasets', there may not be sufficient cells to achieve this metric in the perturbed populations and recommend proceeding with all cells within each perturbed cell population regardless (see sections 2 and 9 of the supplementary information in ref. 15). For the NTC cell population, while Peidli et al. recommend choosing 200 cells, we noted on our KOLF chromatin modifiers dataset that choosing 400 NTC cells yielded a better approximation of asymptotic Edist (Supplementary Fig. 1e).

We performed Edist testing per experimental batch (which we refer to as batch anndata). Within each batch, preprocessing was performed as in scPerturb using scanpy: (1) reads were normalized per cell; (2) genes were filtered that were not detected in at least three cells; (3) reads were log-normalized; (4) the top 2,000 HVGs were selected by flavor = 'seurat_v3' using a class-balanced anndata object where each perturbation had exactly 25 cells; (5) the batch anndata object was subset to these 2,000 HVGs and (6) PCA using 50 components was computed on the batch anndata. Then, a set of 400 core NTC cells were randomly selected and, for each perturbed cell population (using all associated cells, which was at minimum 25), Edist testing versus these NTC cells was performed with 10,000 permutations using the scPerturb Edist method (which we modified to accept a random seed to produce reproducible results). Lastly, to aggregate results across batches, the raw *P* value from each individual test was corrected using Benjamini–Hochberg correction across all perturbations from all batches. Significant perturbations are defined as those having an adjusted *P* value < 0.05 (Extended Data Fig. 2g). To assess the actual Edist (that is, the value on the *x* axis in Extended Data Fig. 2g), we use the value output by the scPerturb Edist method. We further performed an analysis of Edist sensitivity to sample size on our dataset (Supplementary Fig. 1f–j), observing trends similar to those reported by Peidli et al.¹⁵ wherein the Edist is inflated in the context of low cell count, while the *P* values (number of significant perturbations) are correspondingly lower.

To determine DEGs, we leveraged DESeq2 as per the recommendation by Peidli et al.¹⁵. As DESeq2 is inherently a bulk RNA-seq modality, we generated three pseudoreplicates as suggested by Hafemiester and Halbritter¹⁰⁶, using 85% of cells associated with each perturbation to generate pseudoreplicates and choosing at random the same number of NTC cells to form control pseudoreplicates. DESeq2 is then performed, where perturbed cells are compared to NTC cells from the same batch. For clarity, by design, each perturbed gene (and associated sgRNAs) was exclusively present within a single experimental batch, while the same set of NTC sgRNAs were present in every experimental batch. Therefore, a batch term was not supplied in the design matrix for DESeq2. DE was run within each batch independently to extract both the *P* and the adjusted *P* values from DESeq2. Significance was assessed per gene at adjusted *P* < 0.05. Upregulated and downregulated perturbations were determined by the sign of the log₂ fold change reported by

DESeq2 (Extended Data Fig. 2h, left) Note that Extended Data Fig. 2h,k only shows perturbations with >10 DEGs for visualization clarity purposes. To determine a dataset-based threshold for the number of DEGs, we performed the same DE test procedure for each NTC sgRNA versus all other control cells per batch. We found that 95% (FDR of 0.05) of NTC sgRNAs induced less than one DEG (per our testing regime) in batches ALPHA and GAMMA and two DEGs in batch BETA (Extended Data Fig. 2h, right). Perturbations that induced more DEGs than this cutoff within each batch were considered significant. Lastly, we filtered for perturbations that were considered significant from both the Edist test and the DE analysis (Extended Data Fig. 2i).

Perturbed-gene-level Edist and DE testing. We then assessed the impact of perturbation at the gene level by viewing cells that received individual perturbations targeting the same gene as having received the same gene perturbation. The same Edist testing (Extended Data Fig. 2j) and DE analysis (Extended Data Fig. 2k) as mentioned above was performed at the gene level. For the DE FDR, the same batch thresholds were used as at the perturbation level. We hypothesize that repeating this procedure at the gene level helps to correct for some instances of aberrant (that is, nonspecific) sgRNA, as, for a given gene, if two different sgRNAs exhibit drastically inconsistent phenotypes, when combined, they will have a within-population variance that is too large to have a significant Edist test result. A majority of perturbed genes passed the second round of Edist testing (Extended Data Fig. 2j). At this stage, we also considered filtering out genes with only a single associated sgRNA after the perturbation-level filtering. Eventually we decided to retain these genes, at the risk of including some genes that are a result of nonspecific sgRNA binding (but nonetheless do have sufficient on-target efficacy), to avoid removing true signal. For example, UHRF1 (a bona fide interactor of DNMT1) was well resolved in our small-scale screen but only had one sgRNA passing thresholds in the genome-scale screen. Upon further downstream analysis, we were able to recapitulate the DNMT1–UHRF1 relationship (Extended Data Fig. 4b).

Strong perturbation selection and prioritization. To select strong perturbations, we selected those perturbed genes that were significant from both the gene-level Edist and the DE testing. To prioritize strong perturbations, we ranked each strong perturbation by its total number of DEGs and Edist. We then averaged these ranks to arrive at a final ranking (Extended Data Fig. 2l and Fig. 1i). Alternate rankings based solely on Edist or the number of DEGs are in Supplementary Fig. 3b,c.

NTC normalization. After selecting strong perturbations, we sought to represent their transcriptional profile on a scale relative to NTCs using the methodology from Replogle et al.³⁶. Counts per cell were normalized to the median number of UMIs across NTC cells and log-transformed by sc.pp.log1p. Then, expression profiles for cells within a batch were z-normalized to the mean expression profile of NTC cells from the same batch. That is, $n = \frac{x - \mu_{\text{NTC}}}{\sigma_{\text{NTC}}}$, where *n* is the normalized expression profile of a cell, *x* is the unnormalized expression profile of a cell, μ_{NTC} is the mean expression profile of NTC cells and σ_{NTC} is the s.d. of expression profiles of NTC cells. We clipped normalized expression values to be in the range [−10, 10] to handle extreme outliers. We also remark, as was also observed by Replogle et al.³⁶, that this procedure has the additional effect of performing batch correction as NTC cells were treated with the same population of sgRNAs across batches.

Feature gene selection. Feature gene selection was performed using the methodology from Replogle et al.³⁶. In Replogle et al.³⁶, the feature set is the union of (1) top ten DEGs per strong perturbation ranked by the Anderson–Darling test statistic and (2) genes with mean > 0.25 UMIs per cell with variance in the top 30% of the dataset. For (1), as we used DESeq2 instead of AD, we took the top five upregulated

and downregulated DEGs per strong perturbation with absolute $\log_2$ -transformed fold change > 2 . For (2), using the top 30% of high variance genes in our dataset yielded a final set of $> 4,500$ genes (which was almost double the 2,319 genes identified by Replogle et al.³⁶), likely because we included intronic reads yielding a larger total gene set size. To match the same order of genes, we instead used the top 10% of high variance genes, yielding a final union set size of 2,189 feature genes. We note that, while choice of this feature set is beneficial for standardizing features to a fixed length, a key trade-off is potential underrepresentation of strong-perturbation-specific effects (particularly in the case of very strong effects) because of the ten DEGs per strong perturbation cap.

Clustering, visualization and annotation of correlated strong perturbation profiles. To represent strong perturbation effects as a single vector, cells per strong perturbation were mean-pseudobulked, forming the final processed data matrix of 1,665 strong perturbations $\times$ 2,189 feature genes (Fig. 11, left). The pairwise Pearson correlations among strong perturbations (as well as feature genes) were then computed (Fig. 11, middle). To assess CORUM Links (that is, pairs of genes within the same CORUM complex), we followed the same procedure as Replogle et al.³⁶. Using the CORUM 5.1 release¹⁰⁷, we subsetting to all nonhomomeric complexes for which at least 66% percent of complex members were in the strong perturbations (or feature genes). Pearson correlation KDE distributions were then plotted for these CORUM pairs versus all non-CORUM pairs (Fig. 11, right).

To identify clusters of correlated strong perturbations and feature genes, we used the same procedure and parameters as Replogle et al.³⁶. Specifically, the strong perturbations correlation matrix was clustered using HDBSCAN (metric = 'precomputed', min_cluster_size = 4, min_samples = 1, cluster_selection_method = 'eom'), identifying 30 total clusters. Note that HDBSCAN is conservative by design and assigned a majority of genes to no cluster³⁶. For the feature genes, we first z-normalized each feature gene correlation profile to equate the MDE distance optimization problem to a correlation optimization problem. That is, $z = \frac{n - \mu_n}{\sigma_n}$ where z is the 'internally' normalized pseudo bulk expression profile of a perturbation, n is the normalized expression profile of a cell, μ_n is the mean value of the normalized expression profile and σ_n is the s.d. of the normalized expression profile. We then embedded these normalized profiles into 20 dimensions by pymde using the parameters n_neighbors = 7 and repulsive_fraction = 5. Clustering was then performed using HDBSCAN (metric = 'euclidean', min_cluster_size = 10, min_samples = 10 and cluster_selection_method = 'leaf').

To produce the 2D MDE embedding visualization of strong perturbations, we used the same procedure and parameters as Replogle et al.³⁶ with noted exceptions. After standardizing pseudobulked NTC-normalized profiles to equate the MDE distance optimization problem to a correlation optimization problem (as performed for feature genes), we embedded the dataset first using sklearn's SpectralEmbedding into 20 dimensions using affinity = 'nearest_neighbors', eigen_solver = 'arpack'. While Replogle et al. used n_neighbors = 7, we noticed that, in the spectral embedding on our dataset, when using this parameter, some of the strong perturbations mapped to the same embedding. Therefore, we increased n_neighbors to 25, which was able to alleviate this. Using the spectral embedding initialization, we further embedded the dataset into two dimensions using pymde's preserve_neighbors function with n_neighbors = 7. We increased the repulsive_fraction from 5 (as used by Replogle et al.) to 10, as we observed an improved average distortion (0.0761) and residual norm (3.6×10^{-6}) by doing so.

On the 2D MDE embedding, we performed Leiden clustering attempting to maximize congruence with the identified HDBSCAN clusters. To do so, we swept through the number of neighbors (k) used for Leiden clustering and plotted the adjusted Rand index (ARI)

and adjusted mutual information (AMI) between Leiden clusters and HDBSCAN Clusters (Supplementary Fig. 2a). We observed that, while increasing k did continue to increase the ARI, not much improvement in AMI was seen past $k = 7$; thus, we used $k = 7$ neighbors for Leiden clustering through leidenalg. We note that most of the discrepancy observed between the two clustering methods was based on the mitochondrial translation cluster, where it was identified as a single cluster in HDBSCAN but split over multiple clusters in the 2D embedding, generally suborganizing by large and small subunits of the mitochondrial ribosome. ARI and AMI between Leiden and HDBSCAN clustering was 0.35 and 0.71, respectively, at $k = 7$. Excluding HDBSCAN C26 (the largest cluster related to mitochondrial translation), which was potentially split over multiple subclusters (Supplementary Fig. 8), ARI and AMI increased to 0.62 and 0.75, respectively. It is important to note a limitation of this approach wherein tuning 2D Leiden clustering to maximize congruence with HDBSCAN biases the 2D partitioning toward the HDBSCAN solution rather than treating it as an independent clustering; alternative choices of k , resolution or clustering algorithm could yield different cluster boundaries, particularly for large heterogeneous groups such as the mitochondrial translation cluster. Furthermore, as with any 2D embedding, the MDE projection is susceptible to distortion artifacts in which local neighborhoods may not faithfully reflect high-dimensional distances, potentially leading to erroneous node positioning and influencing downstream Leiden cluster assignments. Hence, the MDE should be viewed as a useful and largely faithful visualization tool, while the cluster map represents the undistorted (ground truth) pairwise correlation values.

To annotate HDBSCAN strong perturbation and HDBSCAN feature gene clusters, we first collected all possible annotations from gseapy's enrichr tool using the gene sets ['CORUM', 'KEGG_2021_Human', 'GO_Biological_Process_2025', 'GO_Cellular_Component_2025', 'GO_Molecular_Function_2025', 'Reactome_Pathways_2024'] with the background set as all strong perturbations/feature genes. Furthermore, if any CORUM links were present within the cluster, those were noted. On the basis of these annotations, combined with STRING PPI enrichment analysis, clusters were manually assigned a name that encapsulated enriched terms (Supplementary Tables 5–7). To designate a single cluster name, the annotation provided by STRING's functional annotation module from the web portal (<https://string-db.org/>) was used as the primary annotation. Specifically in the case of C20, no strong automatic enrichment was detected using standard functional annotation databases, perhaps because of a lack of physical PPIs in this cluster or incomplete coverage of pluripotency-related terminology in these resources. However, manual inspection of the genes within this cluster revealed multiple factors with well-established roles in pluripotency and early developmental regulation. These include *POU5F1* (ref. 3), *PTPN1* (ref. 31), *ZNF521* (ref. 32), *PSMD8* (ref. 33) and *PHF20* (ref. 34). Because these genes collectively represent known components of the pluripotency network, we manually designated this cluster as 'pluripotency'. Strong perturbation Leiden clusters were annotated in the same way, along with annotations from any strong perturbation HDBSCAN clusters they contained. To annotate the MDE map, annotations were placed near constituting genes relevant to associated Leiden cluster annotations.

UMAP visualizations. First, the NTC-normalized profiles were scaled and embedded by PCA (sc.pp.pca with default parameters). The top 50 principal components were used as features to initialize a neighborhood graph (sc.pp.neighbors with default parameters) and embedded into 2D by UMAP (sc.tl.umap with scanpy default parameters, min_dist = 0.1 (umap-learn default)). The UMAP embedding density of perturbed cell populations was computed through a kernel density estimate (KDE; sc.tl.embedding_density with default parameters). Embeddings were visualized using sc.pl.umap and sc.pl.embedding_density.

Exploration of alternate methods to compute perturbation correlations. One key challenge for this study (and we believe fundamentally all single-cell studies) is the ability to generate faithful cell state representations or embeddings. While we chose ~2,000 feature genes to get a global picture of phenotypic variation, it is unclear whether this is the optimal feature set to distinguish perturbation-specific effects. Therefore, we investigated alternate strategies to assess phenotypic correlation on the chromatin modifiers screen (Extended Data Fig. 1i). In the top row, we computed the Jaccard index using the set of DEGs per strong perturbation as features (that is, the more DEGs that are shared, the higher the Jaccard index). The leftmost plot represents using all DEGs regardless of directionality, while the middle plot is using only upregulated DEGs and the rightmost plot uses downregulated DEGs. In the bottom row (left), we computed a ‘mismatch’ Jaccard index, which we defined as the number of DEGs upregulated in strong perturbation A and downregulated in strong perturbation B plus downregulated in A and upregulated in B divided by the total union of all DEGs in A and B. Put simply, this metric represents the extent to which two perturbations have opposite phenotypic effects (DEGs moving in one direction for A are going the opposite direction for B). We also defined a similarity score, which is the sum of the Jaccard up and Jaccard down metrics divided by 1 plus the mismatch index (bottom row, middle). Lastly, we computed the pairwise Edist (bottom row right) as per scPerturb. When comparing the separation of non-CORUM to CORUM pairs from these analyses, along with the traditional transcriptional correlative analysis, we observed that the traditional correlative analysis did appear to potentially better resolve CORUM complexes (Supplementary Fig. 2b,c). This comes with the caveat that within this dataset, as all perturbed genes were chromatin modifiers, many of them likely exhibited similar effects; thus, the random baseline may be inflated. Nonetheless, we feel that finding the optimal way to determine the feature set to represent perturbation effects is an open area of study that merits further investigation.

Comparison to Replogle et al. K562 and Feng et al. KOLF Perturb-Seq datasets. To systematically compare our findings to those from the K562 genome-scale Perturb-Seq from Replogle et al.³⁵, we downloaded their normalized anndata object containing bulked profiles per perturbation. Note that both data were internally normalized in the same fashion, setting a z score scale relative to the NTCs. When comparing the two datasets, while a majority of genes were shared, ~3,000 genes in KOLFs and ~1,200 genes in K562 did not overlap (based on ENSEMBL gene ID and gene name; Supplementary Fig. 9a). Furthermore, only 430 strong perturbations overlapped on the basis of the list of gene clusters from the supplementary information in Replogle et al.³⁵ (Supplementary Fig. 9b).

We next sought to assess whether the functional correlations between strong perturbations and feature genes we identified from our dataset could equivalently be found from the K562 dataset. Using the aforementioned K562 anndata, 1,351 of 1,665 strong perturbations and 1,351 of 2,189 feature genes could be identified (Supplementary Fig. 9c,d). To keep data unified, we computed all analyses on this subset of 1,351 perturbations by 1,351 genes for both datasets. Upon clustering the correlations of strong perturbation transcriptional profiles, we were visually able to clearly identify a difference in the association between strong perturbations in the two datasets (Supplementary Fig. 9e). We note that the top left patch that appears to be preserved corresponds to MRPL/S-associated genes. Overall, this illustrates that the relationship between perturbations does vary substantially between the two datasets.

To get a sense of the variability in perturbation response between the two datasets, we computed the correlation of transcriptional profiles between the same perturbation in both datasets (that is, the correlation of gene A in KOLF with gene A in K562; Supplementary Fig. 9f). We formed a Gaussian noise baseline from the KOLF dataset, adding

Gaussian noise with spread equivalent to the s.d. of the dataset, representing the case where the K562 dataset represents a noisy resampling of the KOLF dataset, which would be the case if they represented the same data. Upon exploration of the correlation of transcriptional profiles between the same perturbation across the two datasets, while we did observe that this relationship tended to exceed random chance (where the random baseline was a random sample of K562 correlations), it did not exceed the Gaussian noise baseline, suggesting that perturbation responses are cell type specific. Perturbations with intragene correlations > 0.5 were enriched for mitochondrial transcription and translation machinery (Supplementary Fig. 9g).

To visualize the potential similarities and differences between transcriptional profiles induced within each cell type, we projected the normalized transcriptional profile of each perturbation by UMAP (Supplementary Fig. 9h). As potential inherent NTC variability may differ between the cell lines, the average magnitude of NTC-normalized profiles may vary between the datasets; to account for this, each profile was first centered to 0 and scaled to s.d. units. While a majority of perturbations separated between the two datasets on the basis of cell type, highlighting differential cellular response to the same perturbation, we did find several examples of perturbations colocalizing between the two datasets, such as EIF3K, METTL13/14 and INTS10/15. Notably, RNF7 and POU5F1 colocalized in both K562 and KOLF, albeit in different regions.

Strong perturbation HDBSCAN clusters represent the gene clusters with the highest phenotypic correlation in the KOLF dataset and, although the KOLF and K562 datasets exhibit differential phenotypic response across a majority of perturbations (as visualized from Supplementary Fig. 9g,h), we reasoned that some inter-perturbation relationships should be preserved. When applying the same analysis from Supplementary Fig. 9e to only the strong perturbation HDBSCAN cluster relationships, we did visually observe potentially increased agreement in clustering between the two data (Supplementary Fig. 9i). Note that the top left square generally corresponds to HDBSCAN C26 (mitochondrial translation). Furthermore, the distribution of within-cluster mean correlations for K562 data was significantly ($P < 0.001$, Mann–Whitney U -test) higher than a random baseline (Supplementary Fig. 9j).

To quantify which of these clusters are shared and which are KOLF specific, we computed the distribution of pairwise correlations for each cluster in the KOLF and K562 datasets, along with a random baseline of perturbation correlations from the K562 dataset (Supplementary Fig. 10). Shared clusters were those where the distribution of K562 correlations was not significantly less than KOLF correlations (regardless of significance relative to random baseline, to provide the most stringent criterion for KOLF-specific correlations). KOLF-specific correlations were identified as those where K562 correlations were not significantly greater than random baseline and KOLF correlations were significantly greater than K562 correlations. Perturbations that potentially could be found in K562 are those that were significantly greater than random baseline but significantly less than KOLF correlations. Overall, we resolved several KOLF-specific clusters such as pluripotency-related clusters and the SIN3 complex and identified shared complexes such as the Integrator complex that Replogle et al.³⁵ validated in their study. Notably, despite clusters being shared between the two cell lines, they exhibit differential phenotypic responses (Supplementary Fig. 9g,h) and differential relative magnitude of response (Supplementary Fig. 9k), with pluripotency-specific genes (for example, *NANOG* and *POU3F2*) having the largest difference in relative ranking between K562 and KOLF.

We further sought to compare our dataset to the recent Feng et al.³⁹ dataset, which profiled ~7,200 gene perturbations across 34 hiPS cell lines spanning 26 genetic backgrounds. The two studies adopted complementary designs. Feng et al.³⁸ prioritized breadth across genetic backgrounds at relatively low per-perturbation coverage, whereas

our atlas prioritizes higher per-perturbation depth in a single line (Supplementary Fig. 11a). This added depth allowed us to resolve coordinated perturbation responses in KOLF2.IJ that were not well resolved in the Feng et al.³⁸ dataset, including mitochondrial translation, mitochondrial RNA metabolism and a pluripotency-related module (Supplementary Fig. 11b), with the TFIID/TAF general transcription complex being detectable in both datasets (Supplementary Fig. 11c). Our atlas, thus, provides a deep, comprehensive reference for the gold-standard KOLF2.IJ line, complementing the Feng et al.³⁸ study.

Overall, these analyses reveal that there are both shared and differential strong perturbation clusters and transcriptional response between the Perturb-Seq datasets, underscoring the cell-type-specific nature of perturbation atlases and justifying the need for performing these large-scale genetic screens in each cell type of interest.

Genome-scale Perturb-Seq RNA-editing analysis pipeline

Analyzing a genome-scale screen for A-to-I RNA editing is rather challenging for two reasons: (1) the fundamental experimental setup is not natively friendly to such an analysis and (2) the massive scale of the dataset means all operations must be performed in parallelized, chunked and memory-efficient manners. On point (1), all downstream operations were performed on the BAM output files of 10x cellranger count, which produced a single output file per channel. Each BAM line contained a cell barcode tag that uniquely identified the cell that read was associated with. However, as Perturb-Seq is a multiplexed screen, the cells associated with a single perturbation were typically split over many channels; thus, every ‘mother’ BAM file (of which we had 96 across the three batches) had to be demultiplexed and then remerged into thousands of ‘daughter’ BAM files (of which we had 24,051)—one per perturbation containing all reads of all associated cells. On point (2), our dataset contained ~145 billion total reads. Naively streaming through the reads sequentially and assigning them per perturbation was both time and memory prohibitive; thus, operations had to be chunked and split over multiple compute nodes for parallelization. As this type of analysis was not performed previously at this scale, we created a custom set of scripts to disentangle perturbation-specific effects on A-to-I RNA editing.

First, we filtered our anndata object to all perturbations achieving >30% target knockdown and having at least 25 associated cells (Extended Data Fig. 2f). We further removed perturbed genes without at least two associated sgRNAs. NTC sgRNAs passing our quality filters were also retained. We then generated a tabulated file per channel, noting all contained perturbations and associated cell barcodes, where we modified the cell barcode suffix from -N (which is inserted by cellranger aggregate) to -1 (which is the suffix for cellranger count outputs). We then streamed through the channel's associated BAM file in chunks of 10 million reads. For each read, the script extracts the cell barcode and, if it matches an entry in the perturbation-to-cell mapping, assigns the read to the appropriate perturbation group. Reads are written to perturbation-specific BAM files for that channel, with file handles kept open throughout processing to minimize input–output overhead (corresponding to >250,000 intermediate BAM files). BAM files were then merged, sorted and indexed across channels for each perturbation using SAMtools merge, sort and index, respectively. File suffixes were appended to each merged BAM file to designate chunks of size 100. We then used RNAEditingIndexer⁶⁰ (version 1.1) to compute the AEI for each perturbation within a chunk, using the suffixes to identify chunks. Default parameters were used, with the built-in hg38 human genome reference, Alu annotations and SNPs. AEIs across chunks were finally aggregated into a single tabulated file containing the perturbation and associated AEI.

To analyze the AEI data, we filtered for NTC sgRNAs within the central 80% of AEI distribution of NTC sgRNAs from the same batch (an attempt to remove outlier NTC sgRNAs) and kept only remaining NTC sgRNAs present in all three batches, corresponding to 124

remaining NTC sgRNAs. We then computed the AEI z score for each perturbing sgRNA by subtracting the mean AEI and dividing by the AEI s.d. of NTC sgRNA from the same batch. Gene-level AEI z scores were represented as the mean AEI z score across all associated sgRNA. To assess gene-level significance, a permutation test with 100,000 iterations was used between the raw AEI of sgRNAs associated with each gene and raw AEIs of NTC sgRNAs from each batch. Permutation testing helped to address unbalanced sample sizes between perturbed genes with two or three associated sgRNAs and NTCs with 124 associated sgRNAs. Benjamini–Hochberg multiple-testing correction was then applied.

Reporting summary

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

Data availability

FASTQ files can be found through the Sequence Read Archive under BioProject [PRJNA1173491](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1173491). The anndata file containing the aggregated 10x feature-barcode matrix can be found on Figshare (<https://doi.org/10.25452/figshare.plus.27261219>)¹⁰⁸. An online web browser for the dataset can be found on GitHub (https://y-doctor.github.io/KOLF2.IJ_Perturbation_Cell_Atlas/), with further information on data availability. All other data supporting the findings of this study are available within the article and its Supplementary Information. Source data are provided with this paper.

Code availability

Analysis pipeline code and Jupyter notebooks used for analysis are available through GitHub (https://github.com/y-doctor/KOLF2.IJ_Perturbation_Cell_Atlas/).

References

- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- Sieckmann, K. et al. AdipoQ—a simple, open-source software to quantify adipocyte morphology and function in tissues and in vitro. *Mol. Biol. Cell* **33**, br22 (2022).
- Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Proc. 21st Medical Image Computing and Computer Assisted Intervention* (eds Frangi, A. F. et al.) (Springer, 2018).
- Fossati, A. et al. System-wide profiling of protein complexes via size exclusion chromatography–mass spectrometry (SEC–MS). *Methods Mol. Biol.* **2259**, 269–294 (2021).
- Penny, J., Arefian, M., Schroeder, G. N., Bengoechea, J. A. & Collins, B. C. A gas phase fractionation acquisition scheme integrating ion mobility for rapid diaPASEF library generation. *Proteomics* **23**, e2200038 (2023).
- Nourredine, S. et al. NF45 and NF90 regulate mitotic gene expression by competing with Staufen-mediated mRNA decay. *Cell Rep.* **31**, 107660 (2020).
- Méant, A. et al. Proteomic analysis reveals a role for RSK in p120-catenin phosphorylation and melanoma cell–cell adhesion. *Mol. Cell. Proteomics* **19**, 50–64 (2020).
- Houles, T. et al. CDK12 is hyperactivated and a synthetic-lethal target in BRAF-mutated melanoma. *Nat. Commun.* **13**, 6457 (2022).
- Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).

104. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data* **6**, 1–39 (2012).
105. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
106. Hafemeister, C. & Halbritter, F. Single-cell RNA-seq differential expression tests within a sample should use pseudo-bulk data of pseudo-replicates. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.03.28.534443> (2023).
107. Tsitsiridis, G. et al. CORUM: the comprehensive resource of mammalian protein complexes-2022. *Nucleic Acids Res.* **51**, D539–D545 (2023).
108. Nourredine, S., et al. A genome-scale CRISPRi perturbation atlas of human induced pluripotent stem cells. *figshare* <https://doi.org/10.25452/figshare.plus.27261219> (2026).
109. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation forest. In *Proc. 8th IEEE International Conference on Data Mining* (eds Giannotti, F. et al.) (IEEE, 2008).
110. Tchieu, J. et al. A modular platform for differentiation of human PSCs into all major ectodermal lineages. *Cell Stem Cell* **21**, 399–410 (2017).
111. Loh, K. M. et al. Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* **166**, 451–467 (2016).

Acknowledgements

We would like to thank A. Vellal, M. Tong, J. Rainaldi and A. Cloninger for useful discussions and support throughout all stages of the study. We thank M. Ghassemian, director of the Biomolecular and Proteomics MS Facility (University of California, San Diego), and K. Jepsen, director of Institute for Genomic Medicine Genomics Center (University of California, San Diego).

Author contributions

Conceptualization and design, S.N., Y.D. and P.M. Experiments, S.N., Y.D., A.D., Y.H.L., J.N.H., A.F., B.C., A.S., M.M. and E.P. Computational analyses, Y.D., S.N., B.P., S.S. and J.G. Supervision, J.P., K.O., C.M., E.L., T.I., N.K. and P.M. Writing, S.N., Y.D. and P.M., with input from all authors.

Funding

P.M., N.K., E.L. and T.I. disclose support for the research of this work from a National Institutes of Health (NIH) grant (OT2OD032742). P.M. further discloses support from NIH grants (R01CA310063 and R01HG012351), a California Institute for Regenerative Medicine grant (DISC4-19271), a Department of Defense grant (W81XWH-22-1-0401) and University of California, San Diego institutional funds. E.L. further discloses support from a Rubicon grant from the Dutch Research

Council (019.231EN.013). J.N.H. was supported by a European Molecular Biology Organization postdoctoral fellowship (ALTF 556-2022). This publication includes data generated at the University of California, San Diego Institute for Genomic Medicine Genomics Center using an Illumina NovaSeq 6000 that was purchased with funding from an NIH shared instrumentation grant (S10 OD026929) and the National Cancer Institute (P30CA023100). The remaining authors declare no relevant funding.

Competing interests

P.M. is a scientific cofounder of Shape Therapeutics, Boundless Biosciences, Navega Therapeutics, Pi Bio and Engine Biosciences. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. T.I. is a cofounder, advisor and holder of equity for Data4Cure and Serinus Biosciences and he is an advisor and shareholder for Ideaya BioSciences. The terms of these arrangements have been reviewed and approved by University of California, San Diego in accordance with its conflict of interest policies. The N.K. laboratory has received research support from Vir Biotechnology, F. Hoffmann-La Roche and Rezo Therapeutics. N.K. has a financially compensated consulting agreement with Maze Therapeutics, is the president and is on the board of directors of Rezo Therapeutics and is a shareholder in Tenaya Therapeutics, Maze Therapeutics, Rezo Therapeutics, GENIE Lifesciences and Interline Therapeutics. E.L. is a cofounder of GenBio.ai and an advisor for the Chan-Zuckerberg Initiative Foundation, Element Biosciences, Cartography Biosciences, Pfizer and Pixelgen Technologies. The remaining authors declare no competing interests.

Additional information

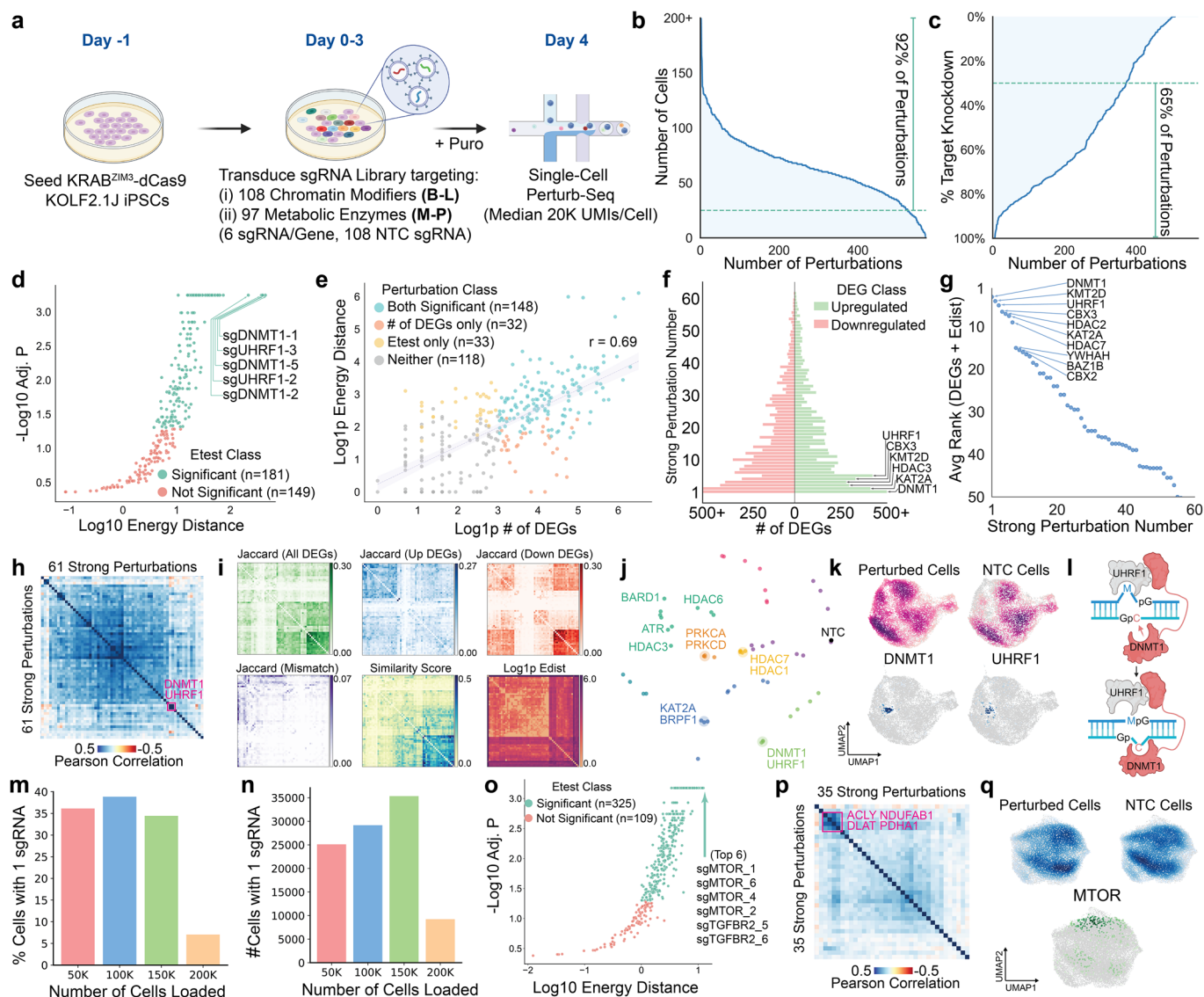
Extended data is available for this paper at <https://doi.org/10.1038/s41587-026-03199-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-026-03199-w>.

Correspondence and requests for materials should be addressed to Prashant Mali.

Peer review information *Nature Biotechnology* thanks Krishanu Saha and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

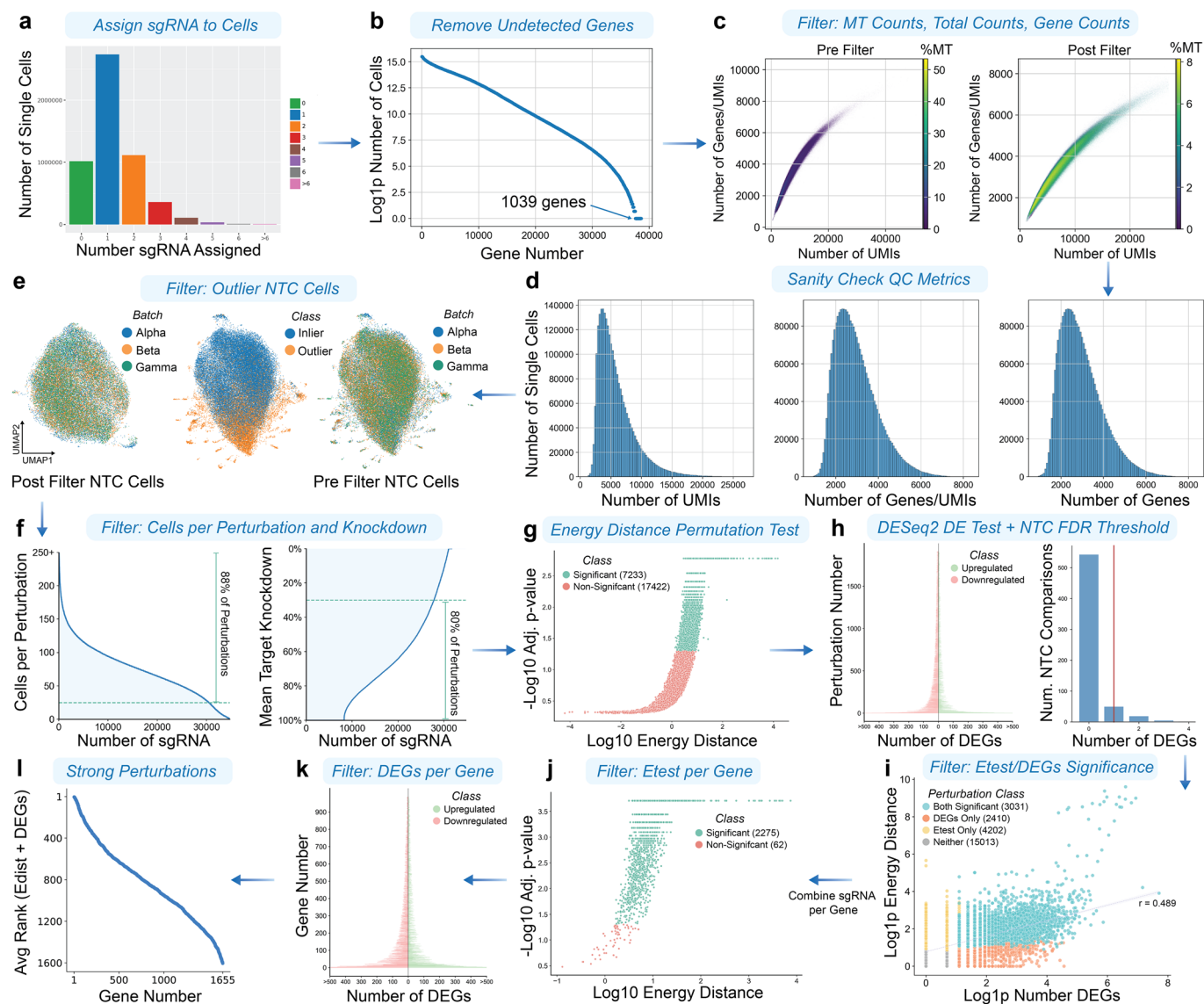
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Perturb-Seq of Chromatin Modifiers and Metabolic Enzymes in KOL2.1J hiPSCs.

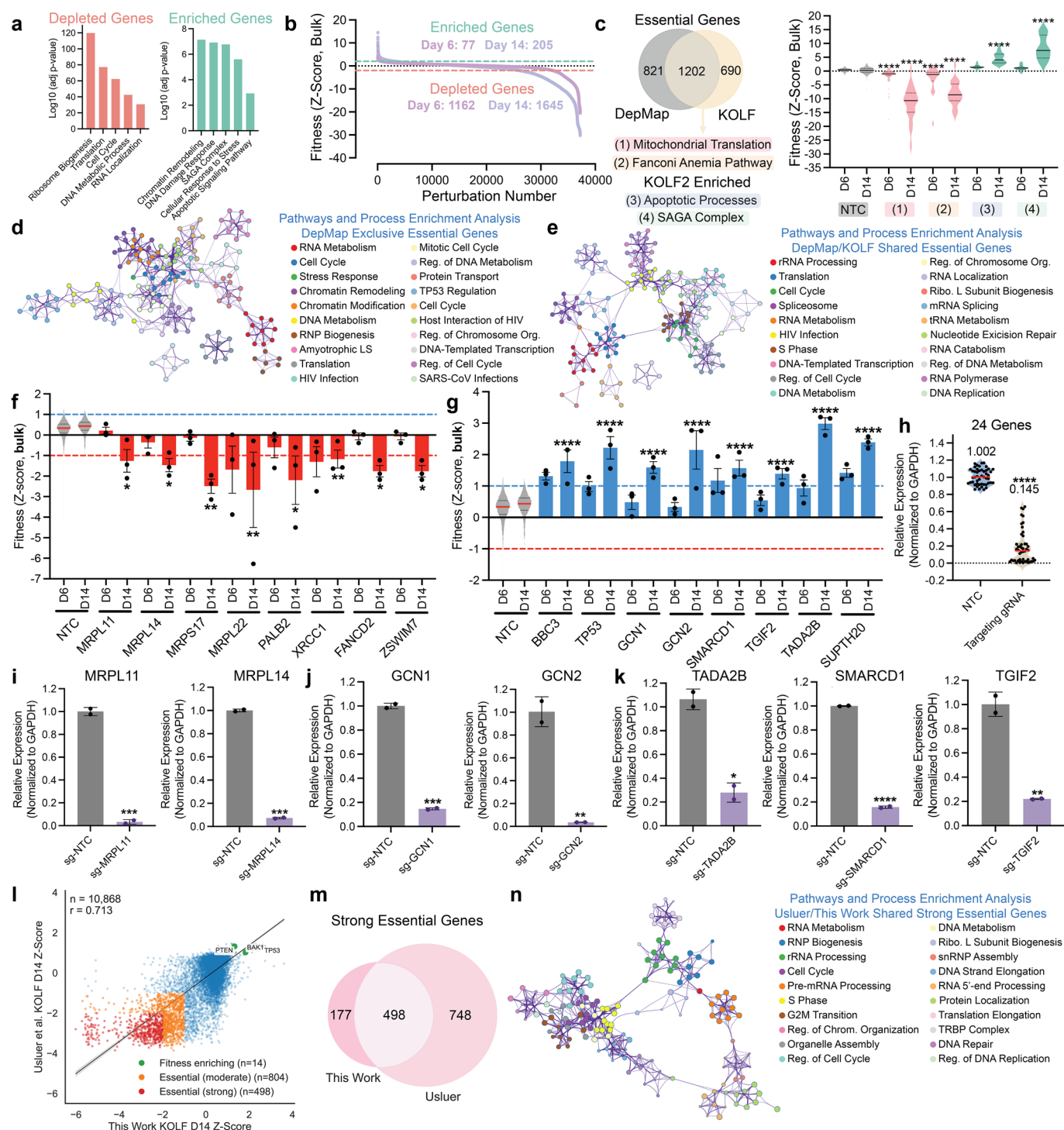
(a) Perturb-Seq pilot screens experimental scheme. KOL2.1J iPSCs constitutively expressing KRAB^{ZIM3}-dCas9 were transduced with a sgRNA library targeting two gene sets: (i) a panel of 108 chromatin modifying genes and (ii) a panel of 97 genes involved in metabolism. 108 non-targeting control (NTC) sgRNA were included. Following 2 day recovery, cells receiving perturbations were selected via puromycin for 2 days. On day 4, cells were harvested for single cell RNA sequencing. (Created using Biorender <https://BioRender.com/f8grmmb>) (b-l) correspond to results from the chromatin modifiers screen, while (m-q) correspond to results from the metabolic enzymes screen. (b) Number of cells per perturbation, indicating that 92% of perturbations (sgRNA) have more than 25 associated cells. (c) On-target knockdown per perturbation, indicating that 65% of perturbations demonstrate repression >30%. (d) Energy distance permutation testing of perturbations vs NTC cells. (e) Relationship between energy distance and number of induced differentially expressed genes (DEGs) per perturbation. Perturbations with significance for both metrics are retained. sgRNA are combined per gene to

form Strong Perturbations. (f) Number of up- and down-regulated DEGs per Strong Perturbation, identified via DESeq2. (g) Waterfall plot ranking Strong Perturbations on their average combined rank of energy distance and DEGs. (h) Pairwise correlation matrix of pseudo bulked transcriptional profiles of strong perturbations. (i) Exploration of alternate methods of clustering perturbations (See Methods). (j) Annotated Minimum Distortion Embedding (MDE) of Strong Perturbation correlation matrix. (k) UMAP density visualization of perturbed and NTC cells (top), DNMT1 and UHRF1 (bottom). (l) Mechanism by which UHRF1 recruits DNMT1 to regulate hemimethylated DNA. Modified from Arita et al.¹⁸ (Created using Biorender <https://BioRender.com/rwvc39u>) (m, n) Effect of loading variable number of cells per 10X channel on yield of cells with 1 sgRNA assigned for metabolic enzymes Perturb-Seq dataset. (o) Energy distance permutation testing of perturbations vs NTC cells. (p) Pairwise correlation matrix of pseudo bulked transcriptional profiles of Strong Perturbations. (q) UMAP density visualization of perturbed and NTC cells (top), MTOR (bottom).



Extended Data Fig. 2 | Data Analysis Pipeline PASTA (Perturbation Atlas Sanitization, Triage, and Analysis). Plots correspond to data from expressed genome wide Perturb-Seq assay. See Methods for full details. Data is first sanitized (a-f), removing outlier cells and non-effective perturbations. Next, perturbations are triaged, removing low-priority perturbations that do not measurably cause cell state alterations (g-k). Finally, analysis is performed to develop a cellular map. (l, Fig. 2). (a) sgRNA are assigned per cell using 10x Genomics guide calling algorithm. Cells without exactly 1 sgRNA are filtered. (b) Genes not detected in any single cells are removed. (c) Cells are filtered based on standard QC metrics proposed by Heumos et al.¹⁰³ (d) Sanity check of standard QC metrics. (e) Filtering of spurious NTC cells. NTC cells are batch-aligned via Harmony¹⁰⁵ and filtered via Isolation Forest¹⁰⁹ assuming a 30%

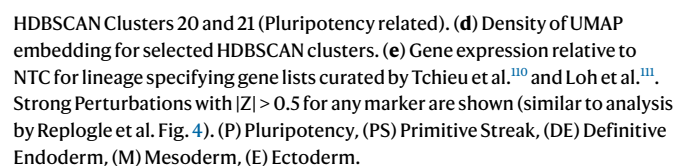
contamination fraction. (f) Perturbations (sgRNA) with less than 25 associated cells and achieving less than 30% on-target knockdown are removed as proposed by Replogle et al.³⁶ (g) Energy distance permutation test vs NTC via scPerturb¹⁵ (h) (Left) Differential expression test of perturbations vs NTC cells via DESeq2. (Right) Distribution of number of DEGs for each NTC sgRNA vs. all other NTC cells is used to set a FDR on the # of DEGs identified. (i) Perturbations with both a significant energy distance test result and more DEGs than the NTC FDR baseline are selected. Perturbations are then combined per targeted gene. (j, k) Energy distance testing and # of DEGs filtering is repeated at the gene level. 'Strong Perturbations' are defined as those which pass all filtering from steps a-k. (l) Strong Perturbations are prioritized based on the average of their energy distance and # of DEGs rank.



Extended Data Fig. 3 | Further Analysis Related to CRISPRi Fitness Screening.

(a) Top 5 GO terms for enriched/depleted perturbations related to Fig. 1(c). (b) Fitness screen analysis Z-normalizing to distribution of NTC sgRNA rather than all perturbations (related to Fig. 1c). (c) Fitness screen analysis Z-normalizing to distribution of NTC sgRNA rather than all perturbations displays the same pathways as when Z-normalizing to all (related to Fig. 1d). (d) Pathways and processes enrichment analysis of DepMap exclusive essential genes (related to Fig. 1d). (e) Pathways and processes enrichment analysis of essential genes shared between DepMap and KOLF2.1Js (related to Fig. 1d). (f, g) Fitness Z-score at Day 6 (D6) and Day 14 (D14) in KOLF2.1J KRAB^{ZIM3}-dCas9 expressing cells at the perturbation level. Each dot represents a different gRNA, 3 gRNA/perturbation and 478 gRNAs for NTC conditions. Red bars are perturbations with z-score < -1 at day 14, blue bars are perturbations with z-score > 1 at day 14. Error Bars

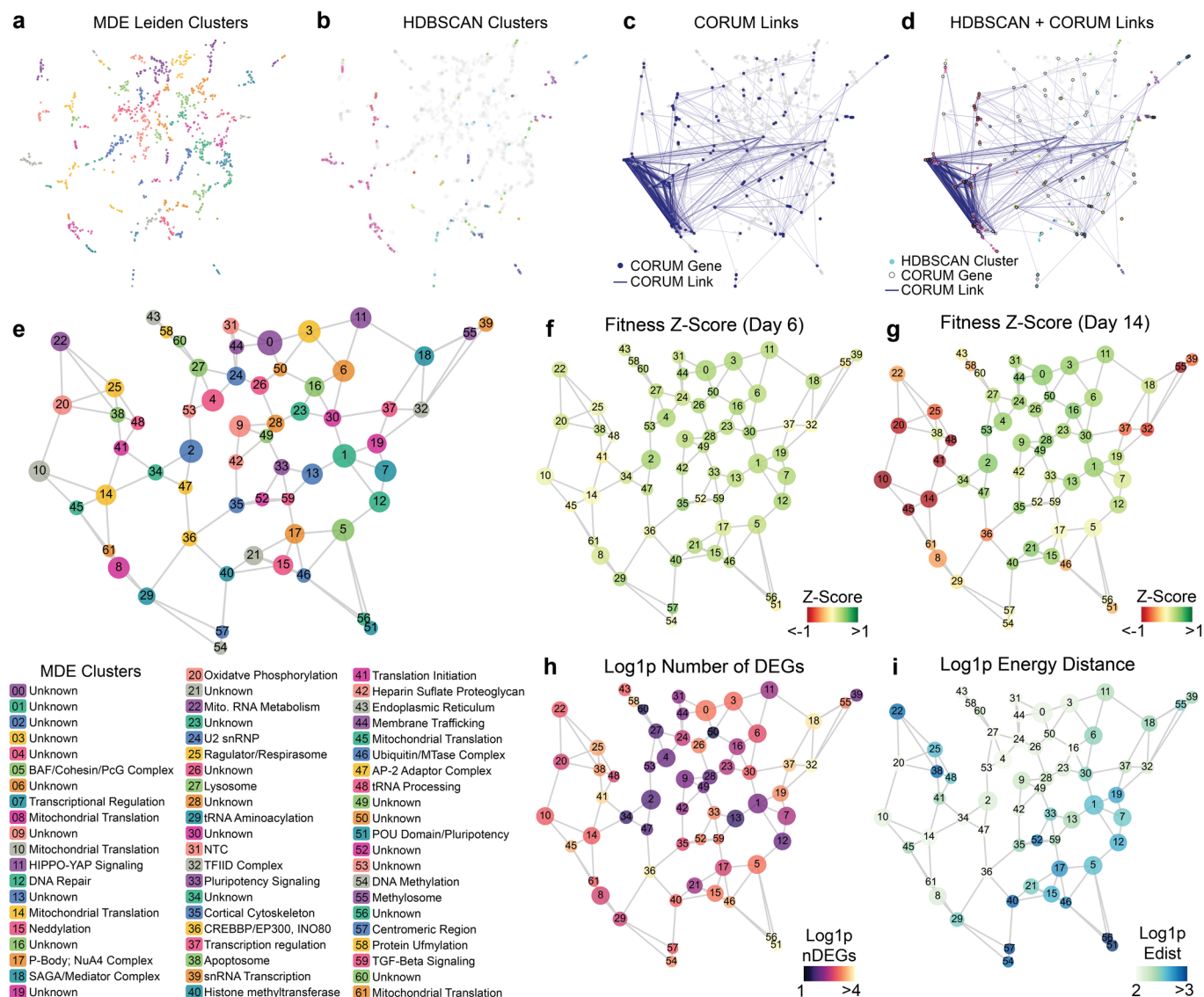
represent SD and asterisks indicate significant differences compared to sg-NTC condition (*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001) (h-k) Gene expression assessed by qPCR between NTC (sg-NTC) and targeting gRNA (sg-Gene) in transduced KRAB^{ZIM3}-dCas9 KOLF2.1J for n = 24 genes (h) including Mitochondrial ribosome large subunit members (i) Stress response genes GCN1/GCN2 (j) and Transcription/Epigenetic Factors (k). Error bars represent SD and asterisks indicate significant differences (*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, n = 2 biological replicates). (l) Correlation of gene-level Z-scores at day 14 between KOLF2.1J hiPSCs from this study and KOLF2.1J hiPSCs from Uslier et al.²⁹). Fitness enriching perturbations are identified by Z > 1 in both datasets, moderate essential Z < -1, strong essential Z < -2. (m) Venn diagram showing overlap between strong essential genes. (n) Pathways and processes enrichment analysis of strong essential genes shared between this work and Uslier et al. dataset.





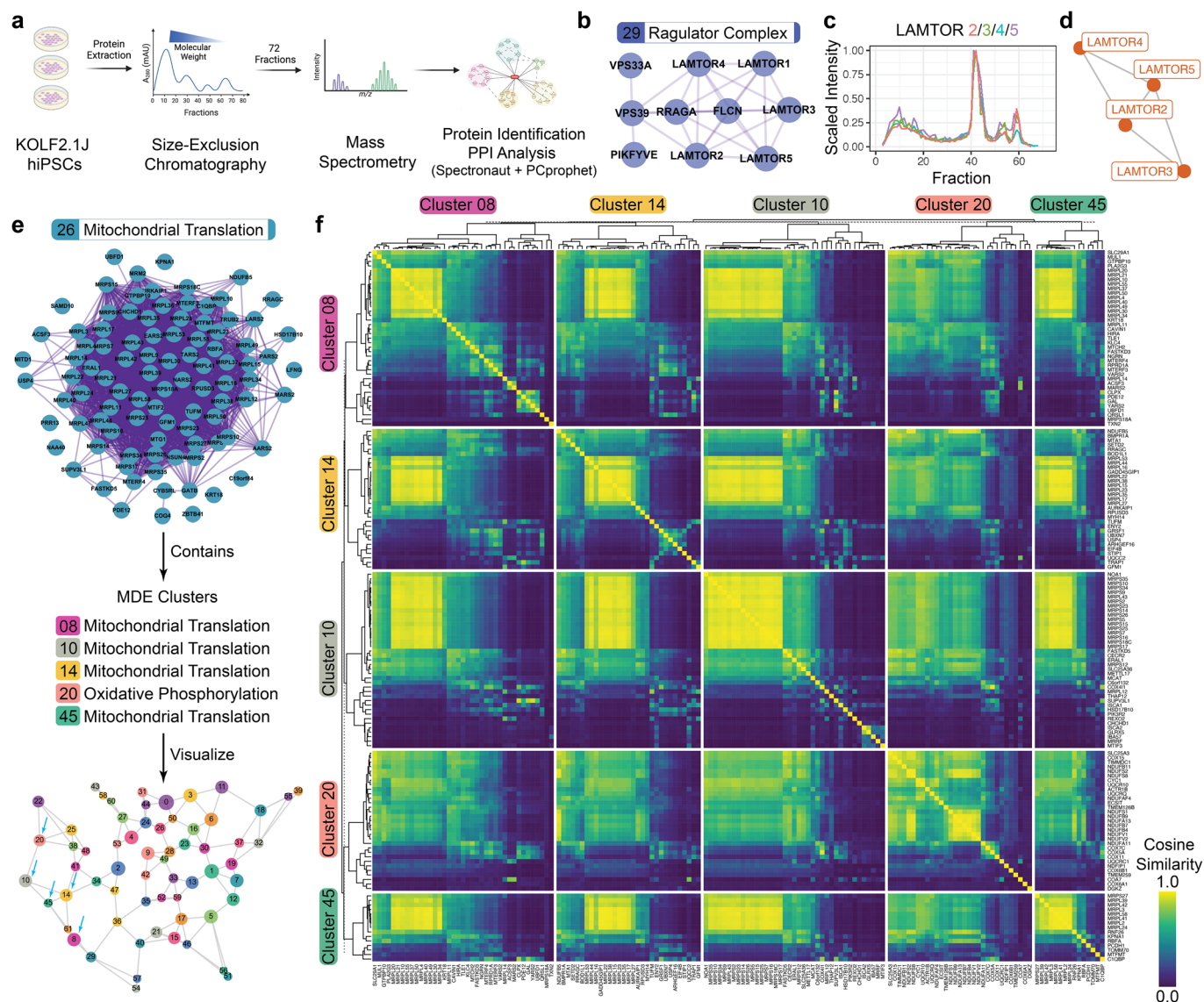
Extended Data Fig. 5 | Summarizing Strong Perturbation and Feature Gene Cluster Relationships. (a) (Left) Pairwise correlation matrix of Feature Genes, subset to those identified in a HDBSCAN cluster. (Right) Annotation of Feature Gene HDBSCAN clusters. (b) STRING PPI networks of selected Feature Gene

HDBSCAN clusters. (c) Summary clustermap of Strong Perturbation and Feature Gene HDBSCAN clusters. Each box (row *i*, column *j*) represents the average Z-score of all Strong Perturbations in column *j* across all Feature Genes in row *i*.



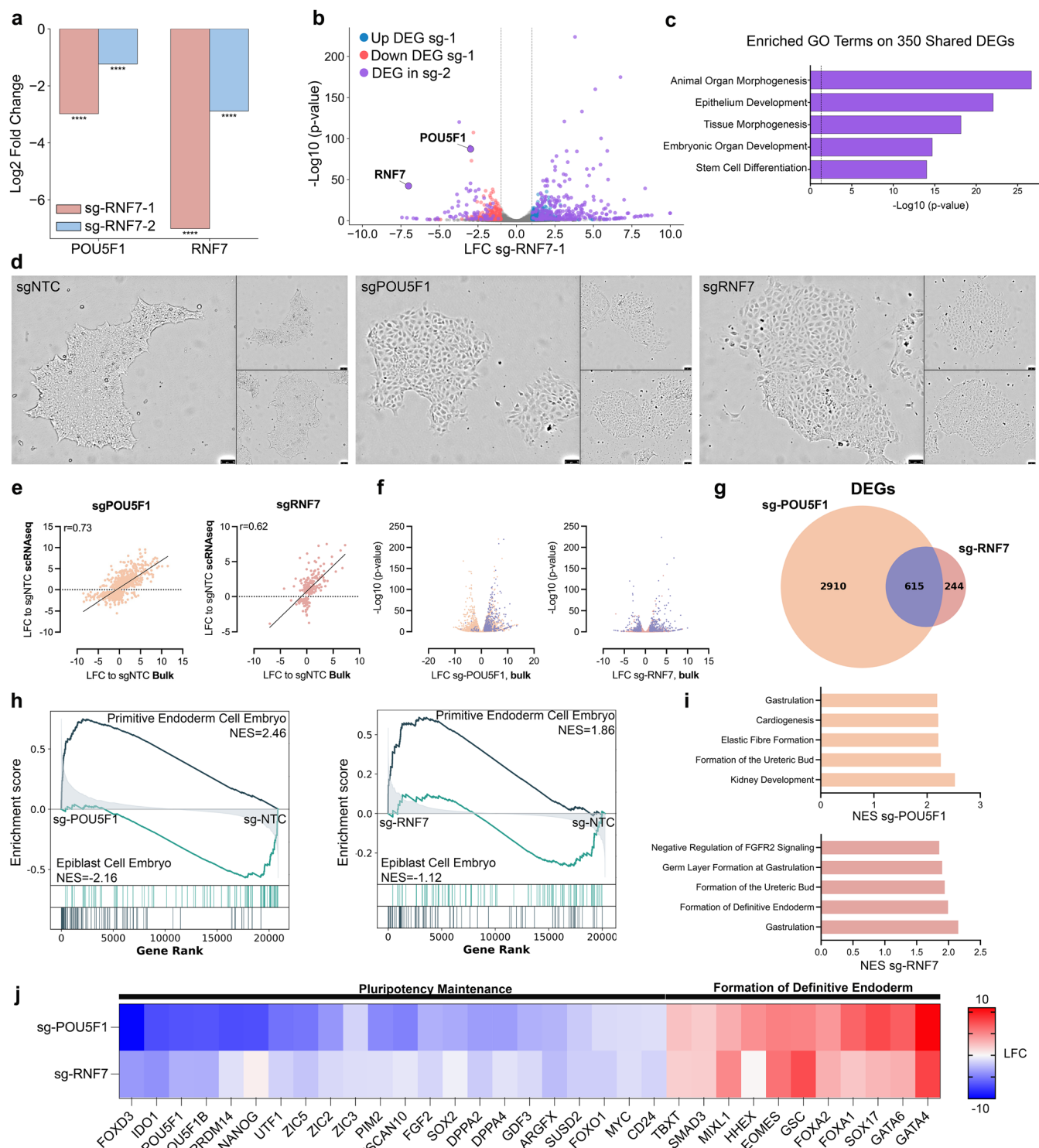
Extended Data Fig. 6 | Visualization of Strong Perturbation Relationships via Minimum Distortion Embedding (MDE). (a) MDE of Strong Perturbation correlation matrix colored by Leiden clusters. (b) MDE map colored by HDBSCAN clusters, where members of the same HDBSCAN cluster are given the same color. Strong Perturbations not assigned a cluster are shown in gray. (c) MDE map colored by genes in CORUM Links (see Methods). Each blue dot is a gene that is a member of a CORUM complex. Blue lines are drawn between pairs of genes

within the same CORUM complex. (d) Overlay of plot (b) on (c). (e) MDE 'skeleton' map of KOLF 2.1J hiPSCs derived from (a). Each node represents Leiden cluster centroids, with the size of the node corresponding to the number of cluster members. Nodes are connected by edges to their 3 nearest neighbors. Nodes are labeled based on enriched GO/CORUM/STRING terms. (f-i) Skeleton MDE nodes colored by median fitness D6 Z-Score (f), median fitness D14 Z-Score (g), median # of DEGs (h), and median energy distance (i).



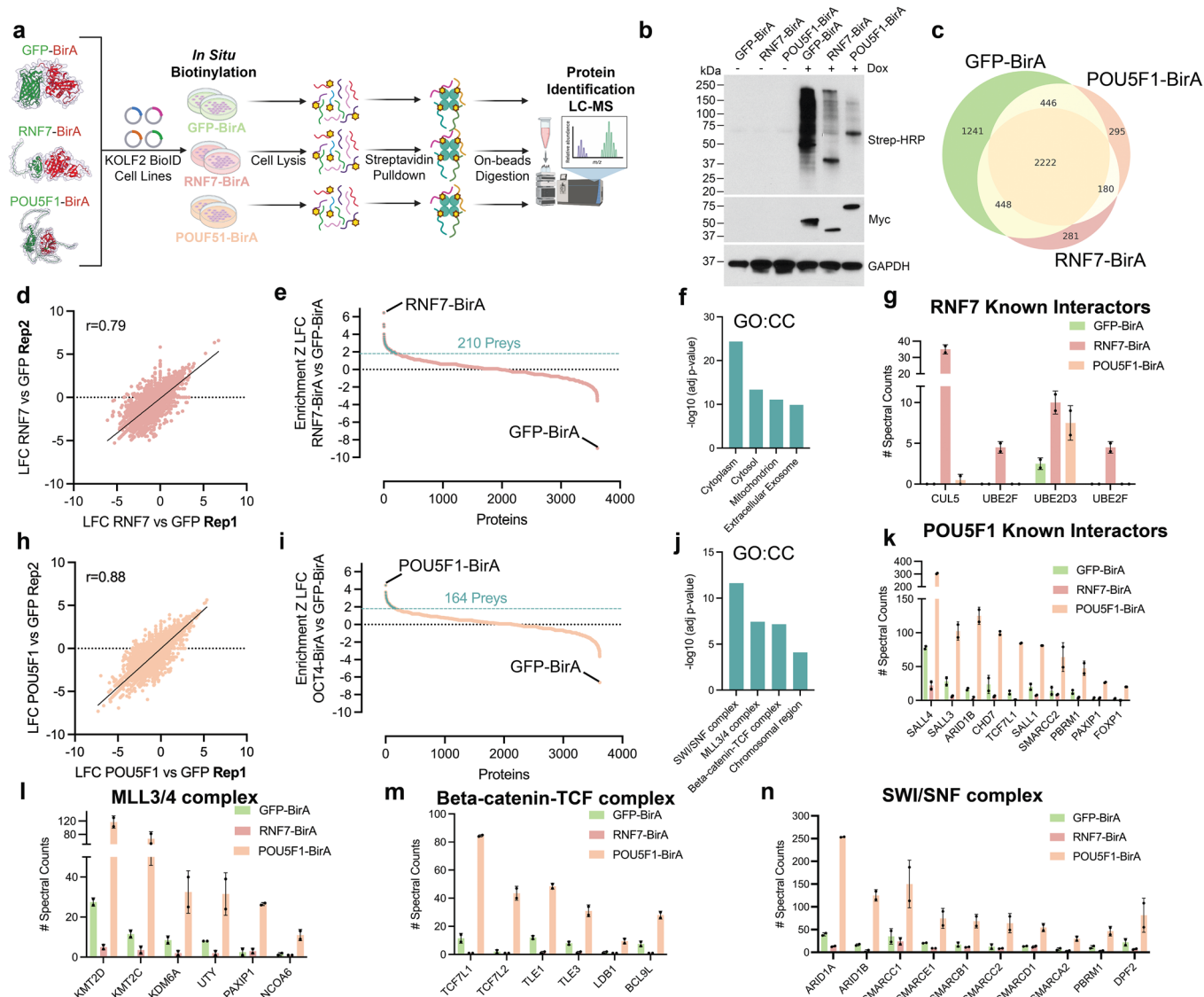
Extended Data Fig. 7 | SEC-MS Validation of LAMTOR and Mitochondrial Translation Complexes in KOLF2.1J hiPSCs. (a) Experimental scheme for size exclusion chromatography coupled with mass spectrometry (SEC-MS) validation experiment. Proteins were extracted from wild-type (WT) KOLF2.1Js and separated by size-exclusion chromatography. Fractions were analyzed by mass spectrometry to identify and quantify proteins by Spectronaut, followed by protein-protein interaction (PPI) analysis yielding interaction networks based on cosine similarity of fractionation profiles. (b) STRING Protein-Protein interaction

network of Strong Perturbation HDBSCAN Cluster 29 (Regulator Complex). (c) Fractionation profile of LAMTOR2/3/4/5. (d) PPI Network predicted from SEC-MS data by PCprophet. (e) STRING Protein-Protein interaction network of Strong Perturbation HDBSCAN Cluster 26 (Mitochondrial Translation). HDBSCAN Cluster 26 contains MDE Clusters 8, 10, 14, 20, and 45 which are indicated on the skeleton MDE with blue arrows. (f) Cosine similarity clustermap of Strong Perturbation fractionation profiles for indicated MDE clusters.



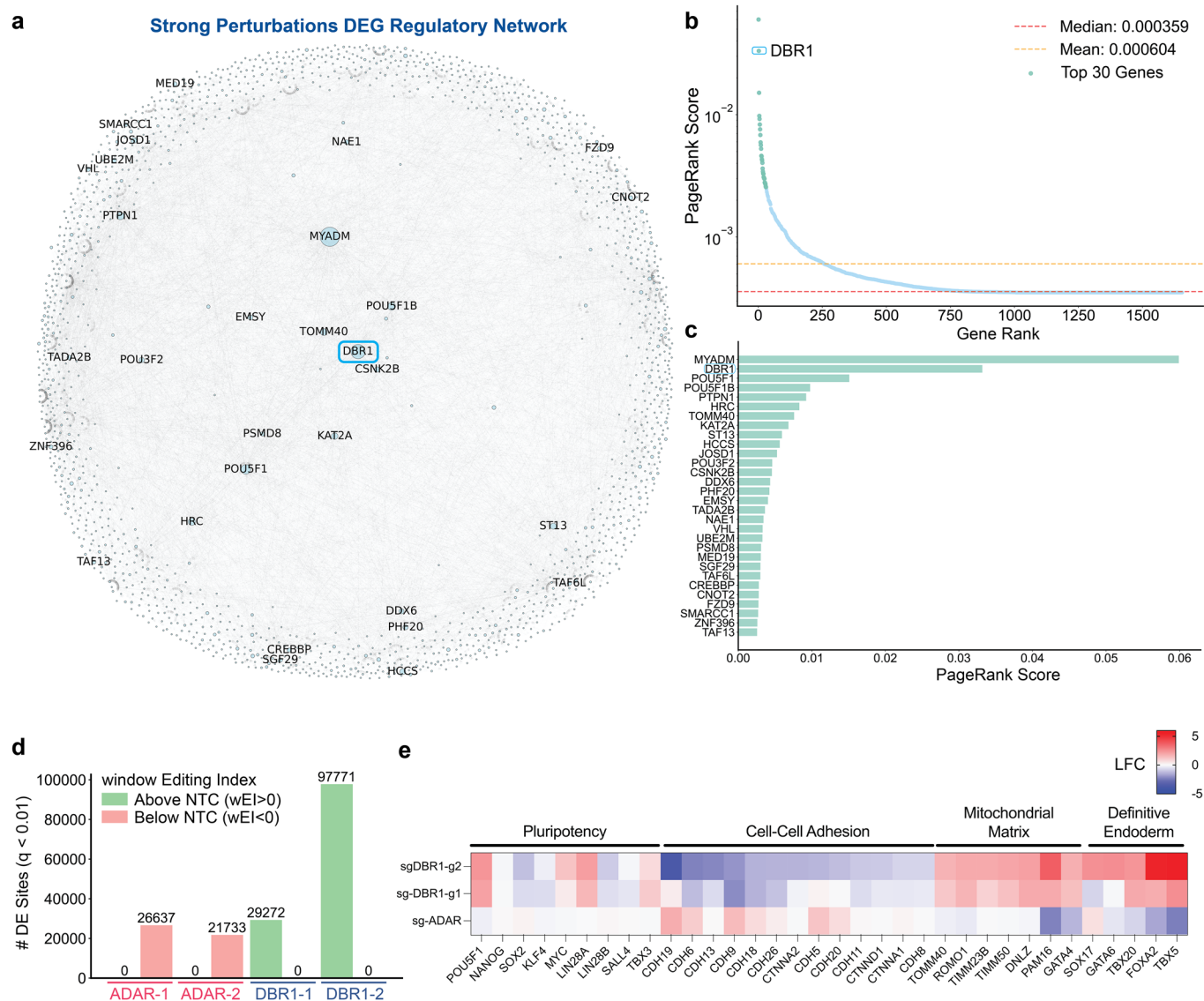
Extended Data Fig. 8 | Analysis of Induced Transcriptional Profiles by POU5F1/RNF7 Knockdown. (a) POU5F1 and RNF7 knockdown induced by two independent RNF7 sgRNA assessed by bulk-RNA-seq (****p < 0.001 per DESeq2). (b) Volcano plot of DEGs (|LFC| > 1, adj. p < 0.05) for RNF7 sgRNA-1, with overlaid DEGs from RNF7 sgRNA-2 (purple), demonstrating that a majority of top DEGs from sgRNA-1 are also identified as DEGs in sgRNA-2. (c) Enriched GO terms (determined via STRINGDB function enrichment tool) from overlap of DEGs between RNF7 sgRNA-1 and sgRNA-2. All further assessments of RNF7 use RNF7 sgRNA-1. (d) Enlarged brightfield images representative of KRABZIM3-dCas9 KOLF2.1J cells with knockdown to RNF7 and POU5F1. Scale bars correspond to 75 microns. Three fields are shown for each condition. (e) Pearson correlation of DEGs identified in Perturb-seq with single-plex bulk RNA-Seq validation for sg-POU5F1 and sg-RNF7 to sg-NTC in KOLF2.1J. (f) Volcano plots of sg-POU5F1

and sg-RNF7 bulk RNA-seq. Purple dots highlight DEGs common to both perturbations. (g) Venn diagram illustrating the overlap of DEGs between sg-POU5F1 and sg-RNF7 knockdowns compared to sg-NTC. (h) Gene Set Enrichment Analysis (GSEA) plots comparing each knockdown condition (sg-POU5F1, left; sg-RNF7, right) against the sg-NTC control, showing enrichment of 'Primitive Endoderm Cell Embryo' and depletion of 'Epiblast Cell Embryo' gene signatures. NES: Normalized Enrichment Score. (i) Bar plots showing the top five significantly enriched Reactome pathways from GSEA, ranked by NES, for sg-POU5F1 (top) and sg-RNF7 (bottom). (j) Heatmap showing expression changes of key genes involved in pluripotency maintenance (left) and definitive endoderm formation (right). Values represent Log₂ fold change from bulk RNA-seq data.



Extended Data Fig. 9 | BioID-Based Mapping of RNF7 and POU5F1 Protein Interactomes in KOLF2.1J cells. (a) Schematic of the BioID workflow. KOLF2 hiPSCs were engineered to express doxycycline-inducible BirA*-tagged constructs (GFP-BirA*, RNF7-BirA*, or POU5F1-BirA*). Following induction and addition of biotin, cells were lysed, biotinylated proteins were captured by streptavidin pulldown, on-bead tryptic digestion performed, and peptides were identified by LC-MS/MS. (Created using BioRender <https://BioRender.com/67kw34i>) (b) Western blot analysis of KOLF2.1J cells expressing GFP-BirA*, RNF7-BirA*, or POU5F1-BirA*, induced for 24 h with 1 μ g/mL doxycycline and 1 μ M biotin. Lysates were probed with streptavidin-HRP to detect biotinylated proteins (top), anti-Myc to confirm BirA* fusion protein expression (middle), and GAPDH as a loading control (bottom). (c) Venn diagram showing overlap of high-confidence prey proteins (FDR $\leq$ 1%) identified in GFP-BirA*, RNF7-BirA*, and POU5F1-BirA* experiments. Numbers indicate unique and shared proteins across the three conditions. (d) Log₂ fold-change (LFC) of spectral counts for RNF7-BirA* versus GFP-BirA* from replicate 1 is plotted against replicate 2. Pearson correlation coefficient ($r = 0.79$). (e) Enrichment Z-score of RNF7-BirA* versus GFP-BirA*;

and 210 significantly enriched preys are indicated (blue dots). (f) Gene Ontology Cellular Component (GO:CC) enrichment for RNF7-BioID 210 preys. (g) Spectral counts of established RNF7 interactors (CUL5, UBE2F, UBE2D3, UBE2F) across GFP-BirA*, RNF7-BirA*, and POU5F1-BirA* experiments. Error bars represent standard deviation of two replicates. (h) Log₂ fold-change (LFC) of spectral counts for POU5F1-BirA* versus GFP-BirA* from replicate 1 is plotted against replicate 2. Pearson correlation coefficient ($r = 0.88$). (i) Enrichment Z-score of POU5F1-BirA* versus GFP-BirA*; 164 significantly enriched preys are indicated (blue dots). (j) Gene Ontology Cellular Component (GO:CC) enrichment for RNF7-BioID 164 preys. (k–n) Comparative spectral count analysis (mean $\pm$ SD, $N = 2$) of GFP-BirA*, RNF7-BirA*, and POU5F1-BirA* BioID datasets for: (k) established POU5F1 interactors SALL4, SALL3, ARID1B, CBX7, TCF7L1, SMARCC1, SMARCA4, PAXIP1 and FOXD1; (l) core MLL3/4 complex subunits KMT2D, KMT2C, KDM6A, UTY, PAXIP1 and NCOA6; (m) β -catenin–TCF transcriptional complex members TCF7L1, TCF7L2, TLE1, TLE3, LDB1 and BCL9; and (n) SWI/SNF chromatin-remodeling complex subunits ARID1A, ARID1B, SMARCC1, SMARCC2, SMARCD1, SMARCD3, SMARCA4, PBRM1 and DPF2.



Extended Data Fig. 10 | Additional Analysis of Regulators of RNA Editing in KOLF2.1J hiPSCs. (a) A graphical representation of the regulatory interactions between Strong Perturbations. Each Strong Perturbation is a node, and a directed outgoing edge is drawn from a node and every other node that is a DEG of that node (that is regulated by it). Node sizes correspond to their PageRank node importance on the reverse graph, and the top 50 Regulators (as identified by pagerank) are highlighted. (b) Waterfall plot of regulators as identified by

PageRank. (c) Barplot of Top 30 regulators as identified by PageRank. (d) Bar plot showing the number of differentially edited RNA sites at a stricter q threshold ($q < 0.01$) identified following knockdown of ADAR and DBR1. Sites with increased editing relative to sg-NTC ($wEI \geq 0$) are shown in green; sites with decreased editing ($wEI < 0$) are shown in red. (e) Heatmap showing expression changes of genes involved in pluripotency, cell-cell adhesion, mitochondrial matrix and definitive endoderm.

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 (<i>n</i>) 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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	NGS Sequencing reads were demultiplexed with bcl2fastq (v2.20) Single-Cell RNA-Sequencing FASTQ files were aligned to Human Genome using CellRanger (v8.0) Mass Spectrometry Proteins were identified via DIA using Spectronaut (v18) for SEC-MS Peaks studio X Imaging Leica STELLARIS8 confocal Fiji package (v2.90)
Data analysis	Analysis was performed using GraphPadPrismv9, CellRangerv8.0, MAGeCKv0.5.8, RNAEditingIndexerv1.1, LoDElv1.0, and custom analysis scripts for the Perturb-Seq analysis which can be found at: https://github.com/y-doctor/KOLF2.1J_Perturbation_Cell_Atlas/

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](#)

The raw .h5ad file for the aggregation of all 10X channels is available at: <https://figshare.com/s/ee85bb1880921326249b>. Raw FASTQs are uploaded to SRA are available at: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1173491>

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

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

The sample size for this study was based on standard biological replicates commonly used in molecular biology analyses. Cell line experiments were performed with a sufficient number of replicates to ensure reproducibility and reliability of the results. Statistical methods to predetermine sample size were not employed, as this study did not involve clinical trials or subjective quantifications that require refined statistical power calculations. Instead, biological replicates were chosen to provide confidence in the observed effects and ensure consistency across independent experiments.

For the single cell Perturb-Seq experiments, sample sizes were chosen to facilitate sufficient per-perturbation coverage in line with previous sample size of n=25 cells per perturbation used by Replogle et al. Cell 2022. n=3 independent sgRNA per perturbed gene were chosen based on results from our pilot perturb-seq experiment using n=6 sgRNA per perturbed gene, wherein analysis demonstrated 3 sgRNA were sufficient to ensure functional perturbation representation (see Methods).

For other experiments, no statistical methods or power analyses were used to predetermine sample sizes. Sample sizes were determined based on historical empirical data, published literature, and established conventions within the field for cell-based molecular biology assays (including quantitative PCR, Western blotting, and immunofluorescence microscopy). These sample sizes are sufficient to capture inherent biological variation within cell culture models and provide adequate resolution to detect robust, reproducible, and biologically significant differences between experimental conditions. The specific number of independent biological replicates for each experiment is explicitly detailed in the respective figure legends.

Data exclusions

No experimental replicates were excluded. For the Perturb-Seq analysis while 6 10x chip channels were removed due to sequencing QC having <5000 median GEX UMIs per cell and <15 median CRISPR UMIs per cell (corresponding to 90/96 retained channels) to ensure sufficient read depth and sgRNA call confidence for all analyzed channels

Replication

Multiple biological replicates were performed for all experiments, including SEC-MS, metabolomics analysis, quantitative PCR, and bulk fitness screens. All attempts at replication were successful, ensuring the reliability and consistency of the results. Sample sizes, including biological n, are clearly stated throughout the manuscript.

Randomization

Randomization was not performed in this study, as no subjective quantification was involved.

Blinding

Blinding was not performed in this study as it does not involve clinical trials or subjective assessments requiring group allocation concealment. All experiments, metabolomics analysis, SEC-MS, BioID, quantitative PCR, and bulk fitness screens, were based on objective measurements and standardized protocols, minimizing potential biases.

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

Methods

- | n/a | Involved in the study |
|-------------------------------------|---|
| ☐ | ☒ Antibodies |
| ☐ | ☒ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

- | n/a | Involved in the study |
|-------------------------------------|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

Antibodies

Antibodies used

Primary Antibodies:

POU5F1/OCT4 – Mouse monoclonal (Cell Signaling Technology, #75463; Millipore Sigma, HPA058267)

Tubulin – Mouse monoclonal (Abcam, ab7291)

RNF7 – Rabbit monoclonal (Abcam, ab181986; EPR12001)

ADAR – Rabbit monoclonal (Cell Signaling Technology, #81284)

DBR1 – Rabbit polyclonal (Thermo Fisher Scientific, PA5-30616)

EIF2A – Rabbit monoclonal (Cell Signaling Technology, #9722)

GAPDH – Rabbit monoclonal (Cell Signaling Technology, #2118)

PKR (Cell Signaling Technology, #3072)

phospho-PKR (Thr451) (ThermoFisher, 44-668G)

Secondary Antibodies:

Goat anti-rabbit IgG, Alexa Fluor 488 conjugate (Invitrogen, A11034; used at 1:800)

Goat anti-mouse IgG1, Alexa Fluor 555 conjugate (Invitrogen, A21127; used at 1:800)

HRP-conjugated goat anti-rabbit IgG (Cell Signaling Technology, #7074)

HRP-conjugated goat anti-mouse IgG (Cell Signaling Technology, #7076)

Validation

POU5F1/OCT4 (Cell Signaling Technology, #75463; Millipore Sigma, HPA058267): Validated by the Human Protein Atlas through immunohistochemistry, western blotting, and protein array analysis, showing consistent performance across applications.

Tubulin (Abcam, ab7291): Validated by the manufacturer using immunohistochemistry, western blotting, and other methods. Cited in over 1,200 publications, demonstrating its reliability and widespread adoption.

RNF7 (Abcam, ab181986; clone EPR12001): Validated by the manufacturer through western blotting and immunocytochemistry. Application-specific validation data is available and widely cited in studies of ubiquitination and neddylation pathways.

ADAR (Cell Signaling Technology, #81284): Validated by the manufacturer using western blotting and immunocytochemistry. Recognizes endogenous ADAR1 p110 and p150 isoforms; used in studies of RNA editing and innate immune signaling.

DBR1 (Thermo Fisher Scientific, PA5-30616): Validated by the manufacturer for western blot applications. Cited in studies focusing on

RNA lariat debranching and intron turnover.

EIF2A (Cell Signaling Technology, #9722): Validated for western blotting. Manufacturer provides validation data in multiple cell lines with expected molecular weight detection; cited in studies of translational control and stress responses.

GAPDH (Cell Signaling Technology, #2118): Extensively validated by the manufacturer for western blotting and immunocytochemistry. Commonly used as a loading control and cited in thousands of publications.

PKR (Cell Signaling Technology, #3072) - Validated by manufacturer by immunohistochemistry and western blotting.

phospho-PKR (Thr451) (ThermoFisher, 44-668G) - Validated by manufacturer by immunohistochemistry and western blotting.

Eukaryotic cell lines

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

Cell line source(s)	HEK293T cell line was acquired from ATCC (CRL-3216) (F Donor) KOLF2.1J iPSC cell line was acquired from Jackson Laboratory (JIPSC001000) (M Donor)
Authentication	All cell lines were purchased from the manufacturer and were not subjected to further authentication
Mycoplasma contamination	Cells were tested frequently for mycoplasma contamination using PCR-based detection, following an adapted protocol from the MycoSensor PCR Assay Kit (Agilent, 302108)
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines used in the study

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>