Cell Genomics, Volume *2*

## Supplemental information

# Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells

**David Wu, Aunoy Poddar, Elpiniki Ninou, Elizabeth Hwang, Mitchel A. Cole, S. John Liu, Max A. Horlbeck, Jin Chen, Joseph M. Replogle, Giovanni A. Carosso, Nicolas W.L. Eng, Jonghoon Chang, Yin Shen, Jonathan S. Weissman, and Daniel A. Lim**

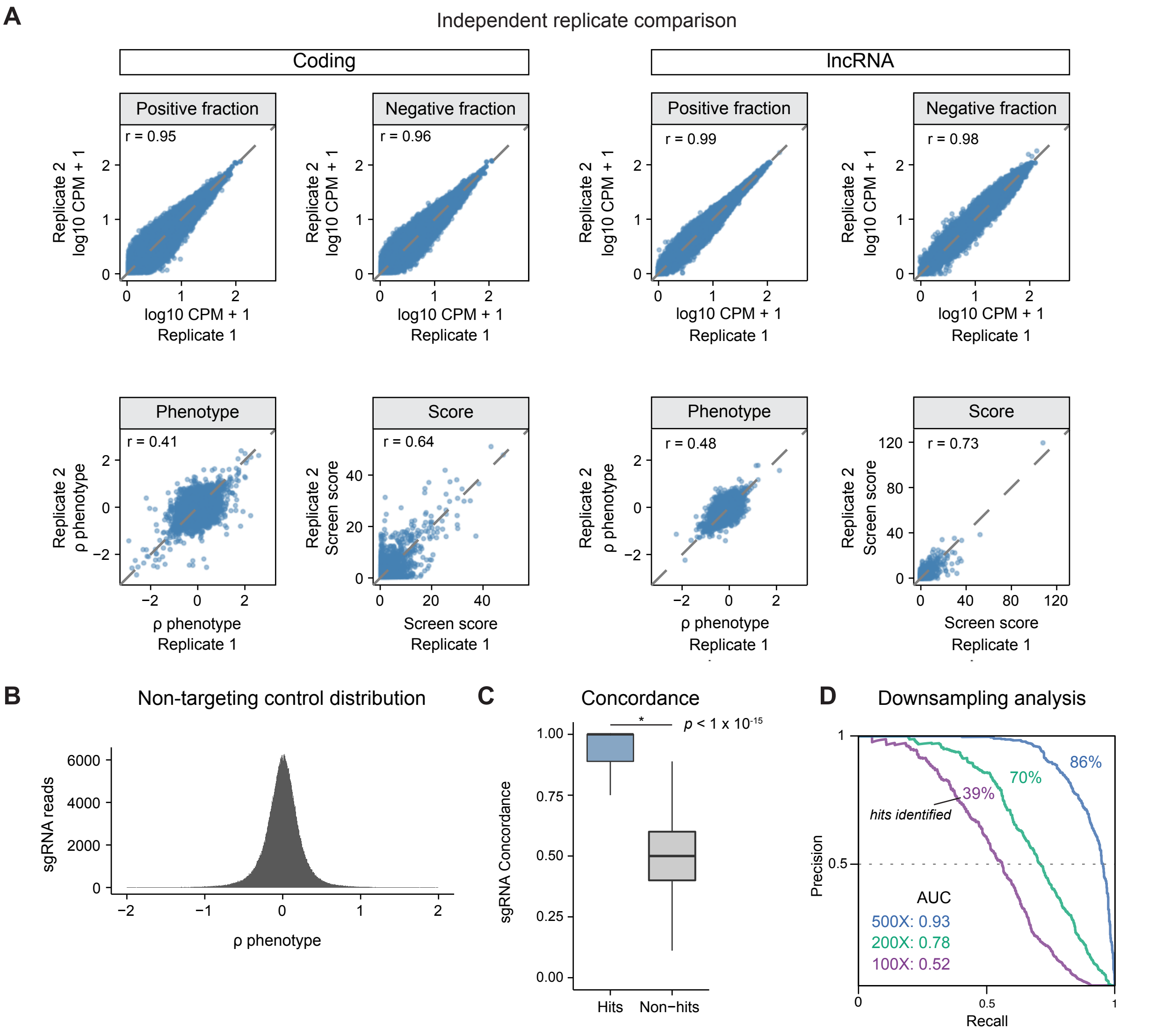


**Figure S1. Genome-wide screen metrics, related to Figure 1**

(A) Replicate analyses of coding and lncRNA screen data show reproducibility between replicates at the level of normalized read counts (CPM), phenotype (ρ), and screen score. The screen score incorporates both effect size and statistical significance and is calculated by taking the product of the magnitude of the ρ *Z*-score and $-\log_{10}$ *p*-value. Pearson correlations (*r*) are reported for each analysis.

(B) Distribution histogram of ρ values for non-targeting controls from genome-wide screen, which were symmetric and centered around zero.

(C) Concordance analysis of sgRNAs. For each hit, the proportion of sgRNAs in the same effect as the hit direction (positive or negative) is calculated; for non-hits (no direction), the proportion of sgRNAs in the positive direction is calculated. Hit sgRNAs were highly concordant and significantly more concordant than non-hit sgRNAs. * *p*-value $< 1 \times 10^{-15}$ as determined by Wilcoxon test.

(D) Precision-recall curves from downsampling analysis to approximate hit identification rates at lower levels of screen sgRNA coverage. Given a starting coverage of >1000X, raw screen data was downsampled to 10%, 20%, and 50% to roughly approximate 100X, 200X, and 500X coverage, represented by purple, green, and blue lines. Precision-recall analysis was performed compared to the full dataset, with median AUC values reported. The percentage of hits identified (relative to the full dataset) are reported for reference.

See also Tables S1 and S2.

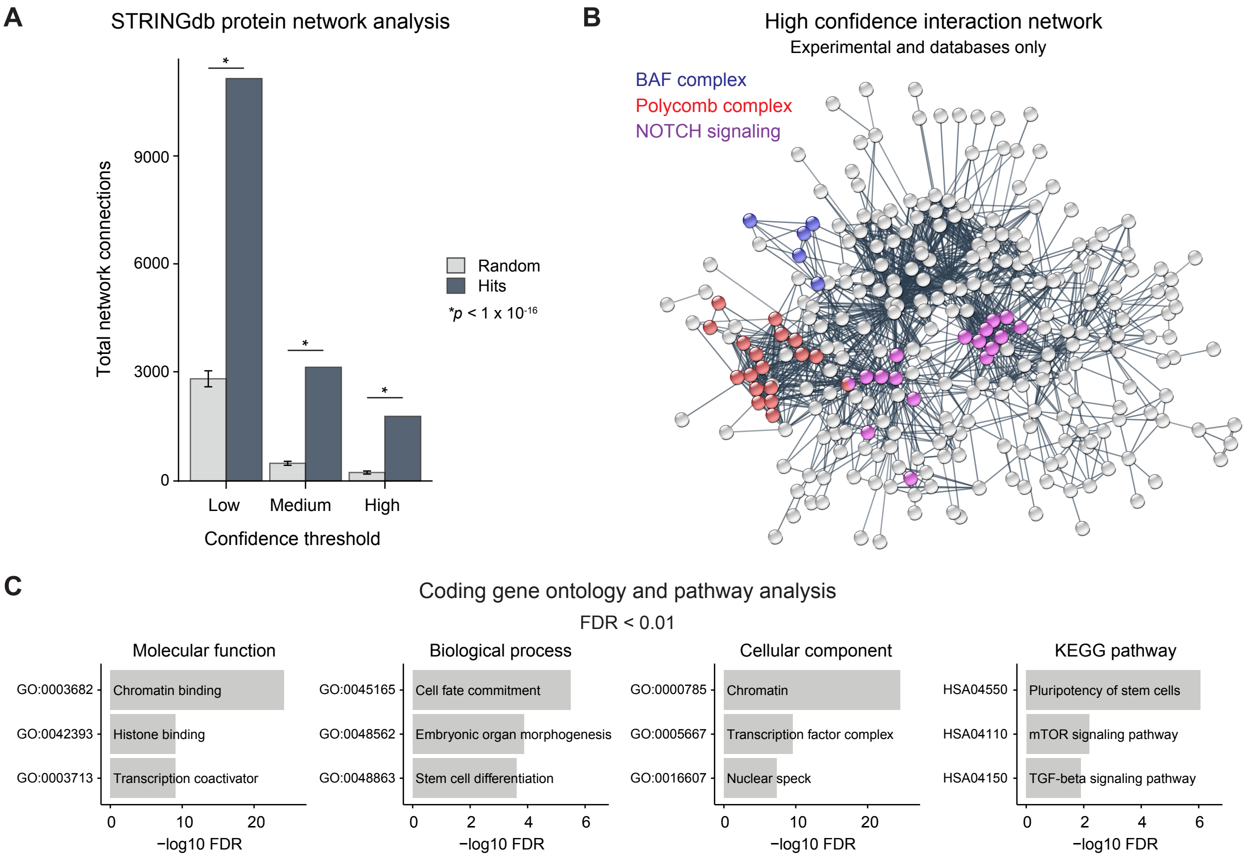


**Figure S2. Bioinformatic validation of genome-wide screen results, related to Figure 2**

(A) Analysis of protein-protein network interactions in the STRINGdb database for screen hits compared to randomly sampled gene sets of equal size. Hits were highly enriched for connections compared to random gene sets at all confidence thresholds. Error bars denote 95% confidence interval from 1,000 random samples. * $p < 1 \times 10^{-16}$ determined from STRINGdb protein-protein interaction enrichment test.

(B) Network graph showing high confidence interactions (from experiments and databases only) between neural induction screen hits, with select complexes and pathways highlighted.

(C) Gene ontology enrichment analysis of neural induction coding gene hits for molecular function, biological process, cellular component, and KEGG pathway terms. All terms were significant at FDR < 0.01.

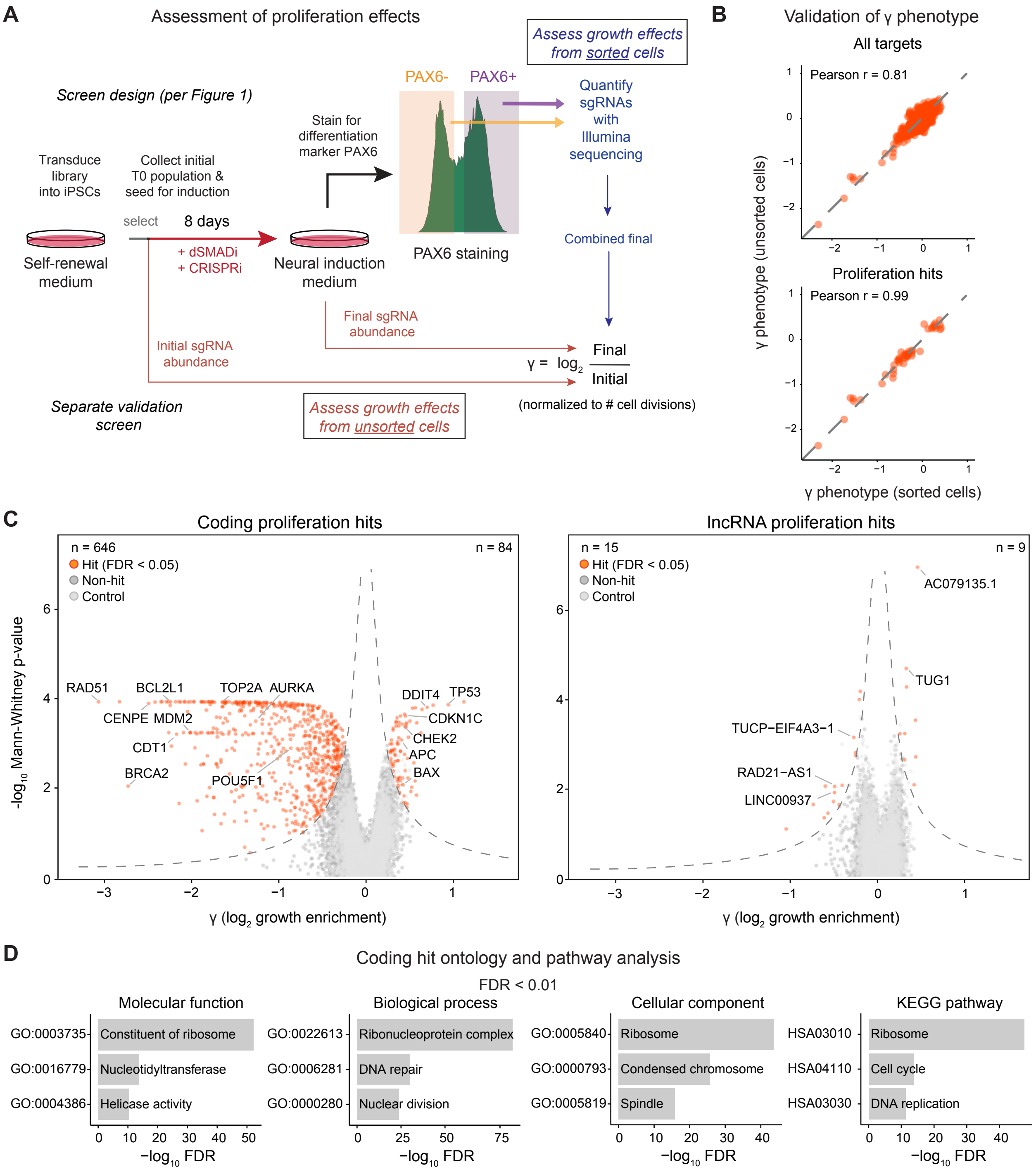


**Figure S3. Genome-wide assessment of proliferation effects during neural induction, related to Figure 3**

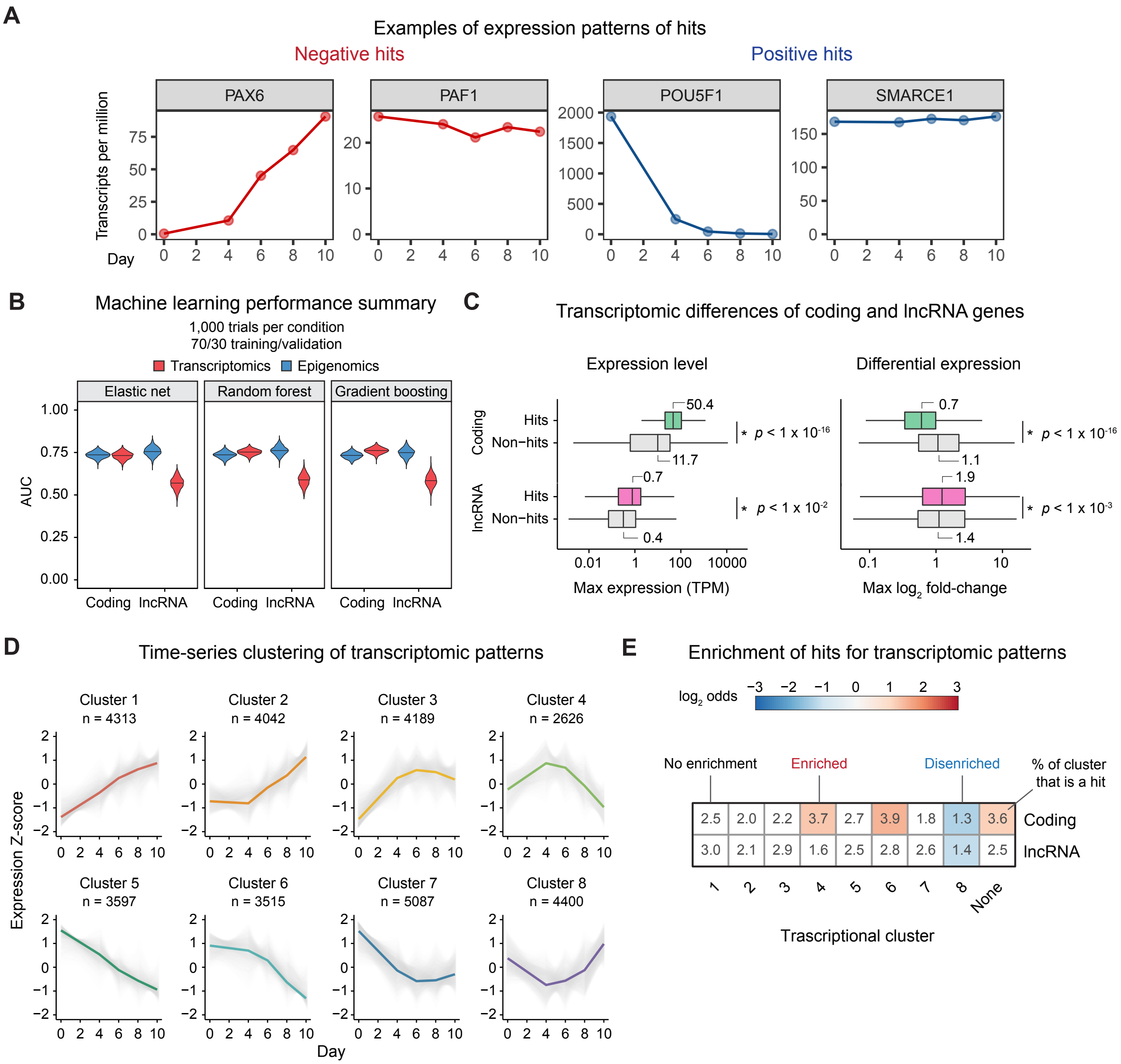
(A) Diagram showing calculation of γ values from sorted cells (primary screen) with day 8 sorted fractions combined into a final time-point to compare to the initial time-point, or unsorted cells (separate validation screen).

(B) Scatter plot showing validation of γ phenotype calculation from the separate validation screen where γ was determined as shown in (A) or by using unsorted cells at the final time-point.

(C) Volcano plots of genome-wide proliferation results for coding genes and lncRNA genes, with X-axis showing screen phenotype γ value and Y-axis showing -$\log_{10}$ *p*-value. Orange dots show hits (FDR < 0.05); dark grey dots show non-hits; light grey dots show non-targeting controls.

(D) Gene ontology enrichment analysis of coding gene proliferation hits for molecular function, biological process, cellular component, and KEGG pathway terms. All terms were significant at FDR < 0.01.

See also Table S2.



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | None |
|---|---|---|---|---|---|---|---|---|---|
| Coding | 2.5 | 2.0 | 2.2 | 3.7 | 2.7 | 3.9 | 1.8 | 1.3 | 3.6 |
| lncRNA | 3.0 | 2.1 | 2.9 | 1.6 | 2.5 | 2.8 | 2.6 | 1.4 | 2.5 |

**Figure S4. Transcriptomic properties of genome-wide screen hits, related to Figure 4**

(A) Temporal expression patterns of individual example hits during neural induction, revealing both high differential expression and stable expression.

(B) Summary plot of AUC distributions obtained from 1,000 trials for each of 3 common machine learning algorithms used to classify hits and non-hits for protein-coding and lncRNA genes.

(C) Boxplots showing the distribution of max expression level (left) and $\log_2$ fold-change (right) of coding genes and lncRNA genes during neural induction. Coding hits in green; lncRNA hits in magenta; all non-hits in grey. * *p*-values determined by the Mann-Whitney test.

(D) Line plots showing 8 significant transcriptional patterns identified from neural induction RNA-Seq time-series clustering using maSigPro, with time-point plotted on the X-axis and expression *Z*-score plotted on the Y-axis. Colored lines represent median for each cluster; light grey lines represent all genes for that cluster.

(E) Heatmap showing enrichment of coding and lncRNA hits in each of the transcriptomic clusters. Numbers represent the percent of genes within the cluster that are hits, and color scale represents the $\log_2$ odds for statistically significant enrichments, determined by Fisher’s exact test (FDR < 0.05).

See also Table S1.

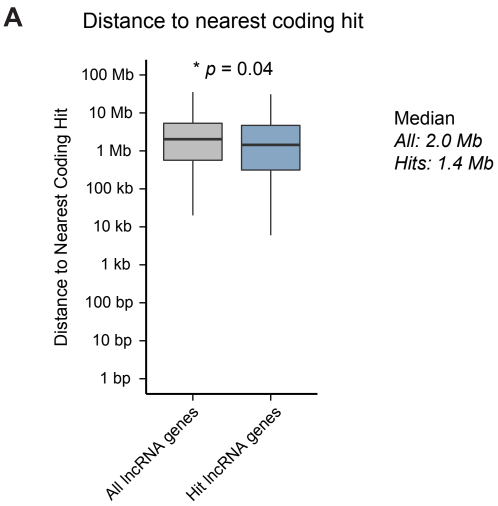


**Figure S5. Coding-lncRNA gene hit distances, related to Figure 5**
(A) Boxplots showing linear genomic distances of all lncRNA genes or hit lncRNA genes to the nearest coding gene hit start site. Hits showed a closer median distance (1.4 Mb) compared to the overall distribution (2.0 Mb), with overall distributions highly overlapping. * *p* = 0.04 as determined by Mann-Whitey test.

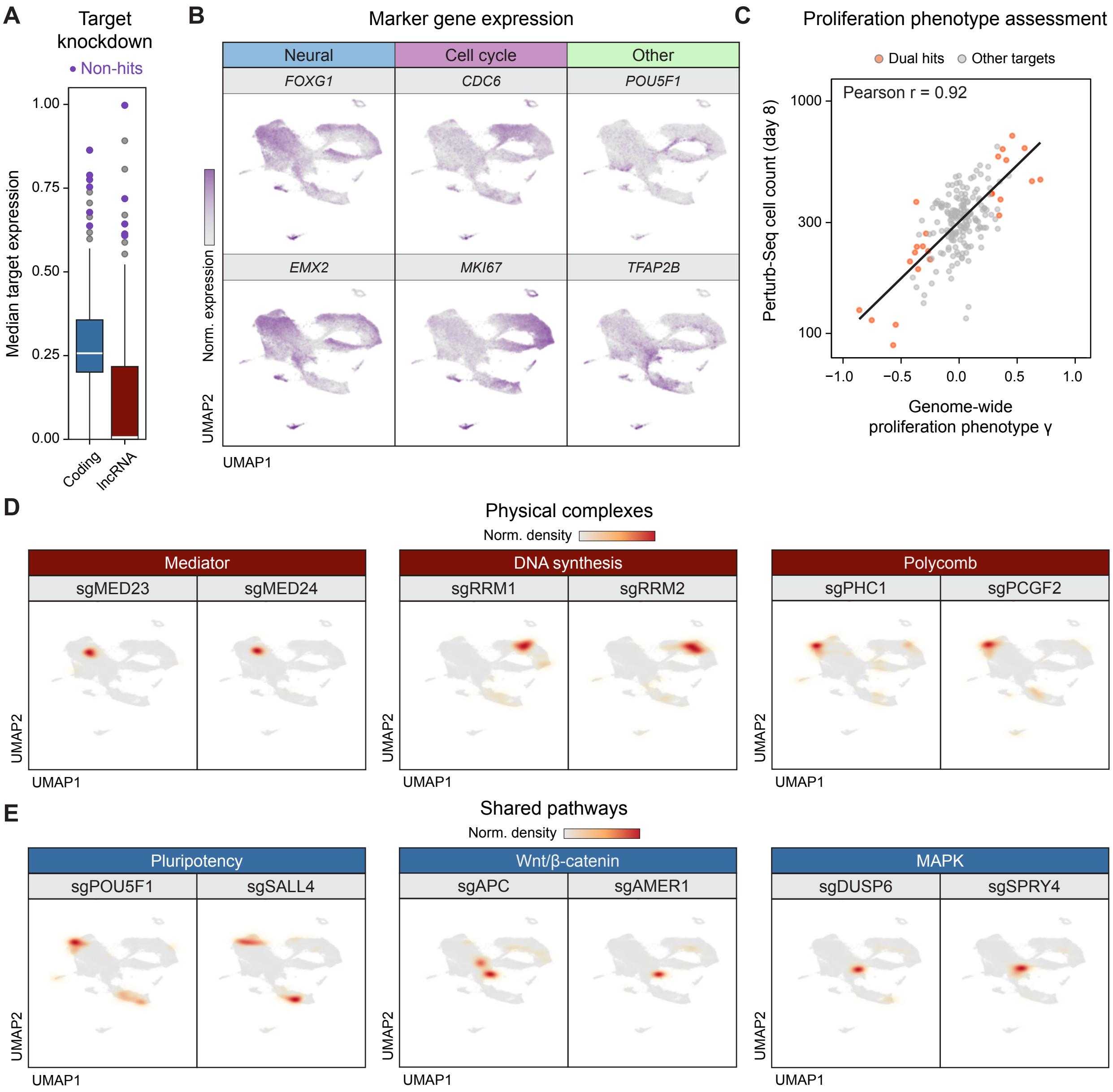


**Figure S6. Functionally-associated targets revealed by Perturb-Seq, related to Figure 6**

(A) Knockdown efficiency of sgRNAs in Perturb-Seq experiment, calculated from median of pseudobulk across six experimental batches. Y-axis shows median target expression (lower expression = higher knockdown). Overall, median knockdown achieved was 80%. Several targets with poor knockdown were non-hits (labeled in purple).

(B) Expression of key markers of neural cells, actively cycling cells, pluripotent cells, and other ectodermal cells, visualized by color intensity on UMAP.

(C) Scatter plot showing validation of proliferation phenotypes by Perturb-Seq, with X-axis showing the genome-wide γ phenotype and the Y-axis showing the day 8 Perturb-Seq 8 cell count. Dual hits (predicted proliferation effects from the genome-wide screens) colored in orange; targets without proliferation phenotypes colored in grey. Pearson r = 0.92 for dual hits.

(D) Normalized density heatmaps of targets with known physical interactions, overlaid on UMAP.

(E) Normalized density heatmaps of targets with known pathway interactions, overlaid on UMAP.

See also Tables S5 and S6.

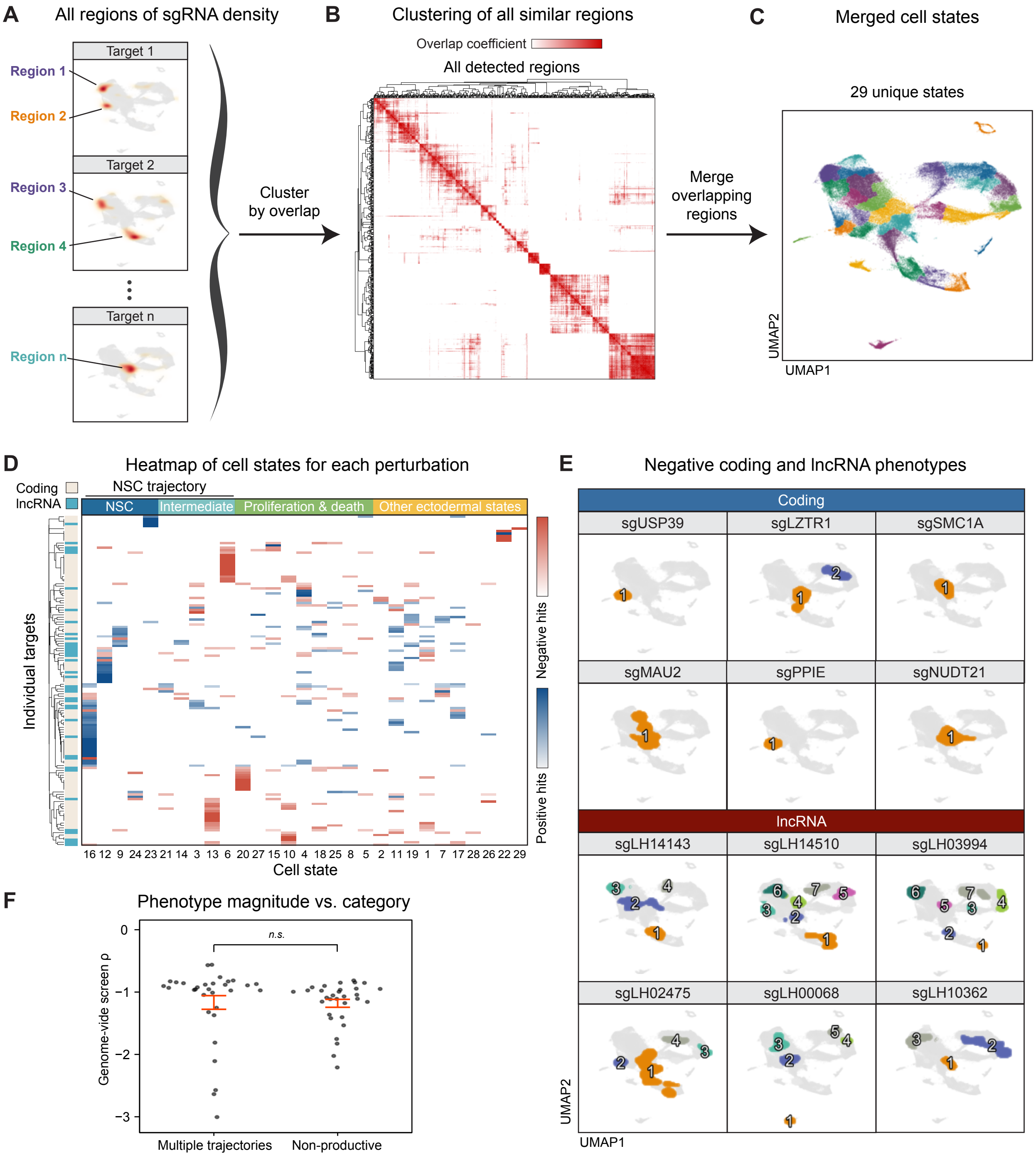


**Figure S7. Perturb-Seq cell state analysis, related to Figure 7**

(A) Overview of regions of high sgRNA density detected across all targets, which were analyzed by overlap similarity for clustering and merging.

(B) Hierarchical clustering of all detected cell states by overlap coefficient.

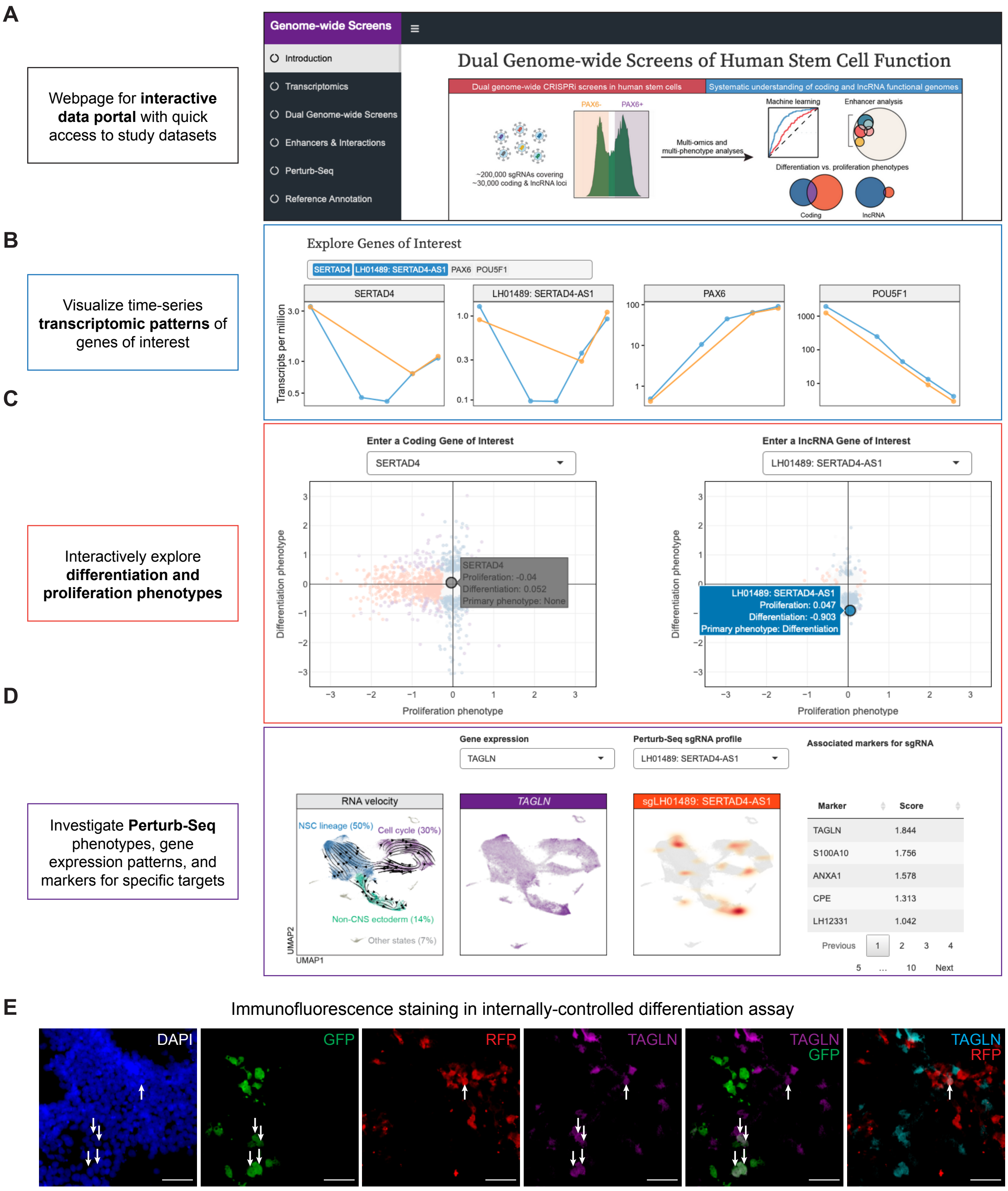
(C) Merging of similar cell states by overlap coefficient into 29 clusters, with k determined by the silhouette method.

(D) Heatmap showing sgRNA density of each target (rows) across the 29 cell states (columns), colored by positive or negative hit direction. Positive and negative hits naturally segregated, although hit direction was not used for clustering.

(E) Examples of phenotypes of negative coding and lncRNA gene hits, with colors representing distinct cell states enriched following sgRNA perturbation (regions detected using DBSCAN).

(F) Genome-wide screen effect magnitude vs. Perturb-Seq phenotype category with the Y-axis showing the neural induction screen ρ. Mean and standard error shown for each group. *n.s.* = non-significant.

See also Table S7.



Webpage for **interactive data portal** with quick access to study datasets

Visualize time-series **transcriptomic patterns** of genes of interest

Interactively explore **differentiation and proliferation phenotypes**

Investigate **Perturb-Seq** phenotypes, gene expression patterns, and markers for specific targets

| Marker | Score |
|---|---|
| TAGLN | 1.844 |
| S100A10 | 1.756 |
| ANXA1 | 1.578 |
| CPE | 1.313 |
| LH12331 | 1.042 |

Immunofluorescence staining in internally-controlled differentiation assay

**Figure S8. Interactive data resource enables intuitive exploration of collective datasets, related to Figure 7**
(A) Interactive data resource webpage (https://danlimlab.shinyapps.io/dualgenomewide) with easy access to datasets and analyses from this study.
(B) Example of exploring the interactive data resource for analyzing genes of interest for their transcriptomic patterns;
(C) Differentiation and proliferation screen phenotypes;
(D) And Perturb-Seq scRNA profiles and suggested markers.
(E) Immunofluorescent staining for TAGLN protein in neurally induced iPSCs sgSERTAD4-AS1 (GFP) or sgControl (RFP). GFP and RFP mark sgRNA+ cells, with arrows indicating sgRNA+ cells expressing TAGLN protein. Downward-pointing arrows indicate GFP and TAGLN double-positive cells; upward-pointing arrows indicate RFP and TAGLN double-positive cells. For each condition, 3 independent replicates were imaged with 10 fields per replicate. Scale bar, 50 µm.