Cell Genomics Transparent Peer Review Record

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

David Wu[1,2,3,4], Aunoy Poddar[1,2,3,4], Elpiniki Ninou[1,2,5], Elizabeth Hwang[3,4], Mitchel A. Cole[1,2,3,4], S.[7] John Liu[1,6], Max A. Horlbeck[7,8,9], Jin Chen[10,11], Joseph M. Replogle[4,12,13,14], Giovanni A. Carosso[1,2], 8 Nicolas W. L. Eng[15], Jonghoon Chang[15], Yin Shen[15,16], Jonathan S. Weissman[7,8,13,14,17], Daniel A. Lim[1,2,18,19,*]

Summary

Initial submission:	Received : May 4, 2022 Scientific editor: Rita Gemayel
First round of review:	Number of reviewers: 4 Revision invited : June 23, 2022 Revision received : July 11, 2022
Second round of review:	Number of reviewers: 4 Accepted : August 22, 2022
Data freely available:	Yes
Code freely available:	Yes

This transparent peer review record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor’s letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Referees' reports, first round of review

Reviewer #1 (Comments to authors)

In the manuscript the authors describe a CRISPRi screen targeting both protein-coding and lncRNA genes, which they term an "ultra-genome-wide" screen. The screen is used to identify lncRNAs with functional roles during differentiation of neuronal cells from embryonic stem cells. The screen provides a nice and important resource of lncRNAs whose perturbation affects the differentiation process and it is followed by follow-up analysis by Perturb-seq and some further description of some of the hits. There are also some general insights about the relative importance of lncRNAs during the differentiation process and reduced importance in proliferating cells. Overall, the paper is of interest to both the lncRNA community and the community interested in CRISPR screens, or more generally, high-throughput approaches for studying lncRNA function. The provided resource should be very useful for the community and the paper sets up high standards for further screen-based studies. As such it should be of interest to the Cell Genomics readership. However, some points need to be addressed before I can recommend publication.

Major comments:
1. Since the main focus of the paper is the new CRISPRi library and the screen, several key technical details should be mentioned/discussed in the Results session: How were the promoters of lncRNAs targeted by CRISPRi selected? (how many per lncRNA? How confident are the coordinates?). How many of the gRNAs were eventually represented enough in the sequencing data to give reliable depletion/enrichment ratios? How many promoters were covered by those? How many lncRNA genes? Can this be used to estimate the minimum number of cells that are required for the screen using the "ultra" library? The authors start from a very large number of cells, it is strictly necessarily retrospectively? Did different gRNAs targeting the same promoter tend to correlate with each other (i.e., if one was depleted was the other depleted as well?). The authors should also provide a BED file (or better, UCSC browser hub) with the coordinates of the lncRNA they use and those of gRNA target regions.
2. It is not clear what the authors define as "dynamically expressed over the time course" (Fig. 1c) and how well the lncRNAs targeted in the screen, are expressed (e.g., in TPM terms) during the time course. Moreover, how often are the genes that appear as targets in the screen are not expressed at all or appear very low during the differentiation time course? E.g., LH09400 which is mentioned as a hit, appears to be expressed at a maximal TPM of 0.11. Can the authors show its RNA-seq coverage? Detect it by some orthogonal methods? Otherwise it seems that the interpretation that many screen hits are just enhancers seems more likely.
3. Is there a tendency for lncRNA screen hits to appear to proximity to coding gene hits? This can also hint at cis-regulatory functions.

Minor comments:
1. I don't think the "ultra-genome-wide" is a suitable term for the screen conducted. It is true that this screen is somewhat more comprehensive than screens targeting just protein-coding or just lncRNA genes, but since screens targeting either have been performed, and in some cells, both protein-coding and lncRNA gene screens have been conducted, and there are still many transcribed regions, such as eRNAs that are not targeted in the screen. A "dual-coding-noncoding" screen or something along these lines would be a more suitable name.
2. Some figures are missing - for example Figure 1D, 2D (in general there discrepancies between the figure numbers and the text).
3. It is not clear what the authors refer to as an "multi-omics" analysis, since only perturbations followed by sorting or RNA expression are used .
4. The "screen score" used in Figure S1 should be clearly defined in the legend.
5. The rationale for doing the analysis at day 8 should be explained.
6. The 32 gRNAs used in validations are targeting how many promoters/genes? 16 genes? Can the 2 gRNAs targeting the same gene be marked in Fig. 2C? That would be more informative than showing the two replicates.
7. Figure 5 can be a supplementary figure, as it does not contain any substantial positive results.
8. The authors should elaborate on the criteria used to select the gRNAs for Perturb-seq, and in particular explain what are the "dual hits".

Reviewer #2 (Comments to authors)

In this manuscript, Wu et al. have performed a large-scale CRISPRi screening targeting both coding and noncoding genes during neural induction from human embryonic stem cells. This led to the identification of hundreds of coding or noncoding genes that are associated with neural differentiation, which were further interrogated by Perturb-seq. The authors have clearly provided a large amount of data for understanding the characteristics of coding and noncoding genes in cellular function. However, there are also some concerns that the authors need to clarify.

Major points:

1. For the Perturb-seq experiments, only 2 independent sgRNAs for each gene were designed. Although the effects of various targets on neural differentiation including the BAF complexes seem to be convincing as different components of the complexes showed similar effects as shown in the UMAP. However, other targets including SERTAD4-AS1, which was selected for further experimental validation, require more sgRNAs to exclude the off-target effects and obtain a more solid conclusion. In this regard, the authors may also utilize "lncRNA (i.e., SERTAD4-AS1) promoter elimination" or "polyA termination signal" strategies to further validate the function of SERTAD4-AS1.
2. The authors applied iPSCs rather than human ESCs for the neuronal induction for both CRISPRi screening and Perturb-seq. This raises a concern that the epigenetic memory of iPSCs may predispose iPSCs to specific differentiation related to the parental cell type used for the pluripotency induction. The conclusion drawn by neural induction from iPSCs (especially for some epigenetic regulators related to epigenetic memory in such context) may not be generic for other human pluripotent stem cells. This should be validated using other human pluripotent stem cells or at least be discussed in the discussion part.
3. Based on the observation that epigenomic data can predict both coding and noncoding hits with similar performance. The authors hypothesized that these lncRNA hits are potential enhancer RNAs. However, further analyses showed that most of these lncRNAs are not enhancer RNAs. What features are these lncRNA hits and what transcripts (lincRNA or others) are these? The authors raised the question without providing any clue in the main text (Page 9-10).
4. In Fig 2C, what are replicate 1 and 2? Are they 2 independent sgRNAs for one specific target? If so, authors should include the information that the 2 sgRNA replicates for one specific gene showed similar effects in the screening?
5. One advantage of Perturb-seq using CRISPRi is that you can detect both the sgRNA and the expression level of its target in the scRNA sequencing. The authors should show some information in the Perturb-seq whether specific target is efficiently silenced when the sgRNAs are simultaneously expressed, especially for SERTAD4-AS1. This will strengthen your conclusion from the perspective of the screening reliability.
6. I would recommend adding neural induction in the title, as the screening was done in the neural induction and not merely in human pluripotent stem cells.

Minor points:

1. "Figure 1D-E" is missing.
2. "Figure 2D" is mislabeled at last line of the part of "Validation of screen results", should be "Figure 2C".
3. "POU5F1" was written wrong at the Figure S3C.
4. "Figure S1B" is mislabeled at third line of the part of "Distinct transcriptomics and epigenomics of coding and lncRNA gene hits"; "Figure 4G" is wrong at fourth line from last. "Figure S1B" is nonexistent.
5. "Figure 6D" is mislabeled at last line of the second paragraph of the part of "Coding gene repression stalls or aborts differentiation, while lncRNA gene repression permits a greater diversity of cell states." "Figure 6D" is mislabeled at the second line from last of the fourth paragraph, should be "Figure 7D".

Reviewer #3 (Comments to authors)

In the manuscript titled with "Ultra-genome-wide screen of coding and lncRNA genome function in human pluripotent stem cells", Wu et al. performed a systematic screening targeting around 30,000 coding/noncoding loci based on CRISPRi and identified hundreds of coding and lncRNA genes functional in human neural induction. Very interestingly, they found lncRNA genes are more prone in regulating differentiation rather than proliferation. Next, they utilized Perturb-Seq to further explore these hundreds of primary hits in neural induction. In general, a serial of extensive screening/analysis in this study provided an informative resource about regulators of neural lineage, and more insights about unique function of lncRNAs in lineage differentiation, which should be a strong candidate to publish in Cell Genomics. However, this manuscript was organized very badly and even unbelievably lacked several pieces of figures (i.e., Figure 1D/E and Figure 2D).

More concerns include:

The major concern is the quantitative readout of screening. Since the Figure 1 failed to match with the manuscript, the reviewer did not understand how the authors calculate the rho value (PAX6 staining) for each sgRNA/cell (Figure 1B-C). Was this information obtained from the screen or validation?

Figure 2D was missing, or should be Figure 2C?

Figure 3 was of interest. A couple of experimental validation could make the finding more obvious and solid.

Figure 8 was not that solid and enough to claim that SERTAD4-AS1 functions in neural differentiation. CRISPRi-based knockdown is difficult to distinct the effect on SEARTAD4 and SERTAD4-AS1.

Reviewer #4 (Comments to authors)

Wu et al developed a novel genome wide screen to determine the coding and noncoding factors that regulate neural induction of iPSC using a large scale CRISPRi technique. The data obtained from this study are novel and informative. It will benefit to our understanding on neural differentiation processes. However, the analysis and validation steps need to be improved substantially. I have following comments:
1. First of all, it is unclear that how rho of 4,523 sgRNAs of non-targeting controls is distributed between PAX6- and PAX6+ population. This is a key information to evaluate the effects of targeting sgRNAs. Because a normal neural induction of iPSC will drive cells to much more toward PAX6+ cells than PAX6- cells, it is not clear if the assumption of "Differentiation phenotypes (rho) were calculated by taking the log2 enrichment ratio of each sgRNA in the PAX6+ versus PAX6- sorted fractions, providing a symmetric measure of the impact on neural induction (as read out by PAX6 protein) on a log-scale" is reasonable or not. This is important as it will affect all downstream analysis and interpretations.
2. Although it is interesting to see the PAX6+ population is altered with certain sgRNAs that target designed coding and noncoding targets, it would be more informative to also analyze the top off-target effects of these sgRNAs to ensure the effects of these sgRNAs were from "authentic" target genes instead of "off-target" genes. The GO analysis results (Suppl 3D) indicates that the specificity of the genes identified is not very informative.
3. The experimental validation of the screen is somewhat circling in logic because the same procedure as screening protocol was used. It might be more informative to use several NPC markers other than PAX6 for the validation step.
4. The biological hypothesis and assumption of "Distinct transcriptomics and epigenomics of coding and lncRNA gene hits" are unclear. Did the authors assume the hit genes will express higher in the late stage of neural induction? What does " individual transcriptomic features" mean? i.e. fold changes? Please formulate these features clearly.
Did the hit lncRNAs will near the individual epigenomic features? What does "individual epigenomic features" mean? i.e. H3K4me3 sites at different differentiation stages? Please formulate these features clearly.
What are the biological interpretation of significant AUC values?
5. "The Perturb-Seq library consisted of 480 sgRNAs for 240 unique targets". Please explain how did these targets selected? why were only these sgRNAs were selected among all hits?
6. "We obtained a total of 78,393 cells that harbored single sgRNA perturbations". Were any cells without sgRNA obtained from single cell RNAseq? How were the "three major cellular trajectories" determined? i.e. based on the cell population without sgRNA perturbation or all scRNAseq results? if sgRNA perturbation data used, if these data used, will it bias the downstream analysis?
7. "The group of non-hit sgRNAs were not statistically distinguishable from controls" It is not clear what were "controls" used here?

8. "stalled and apoptotic phenotypes (collectively "non-productive") represented the majority (56%) of negative coding hits" What are the explanations for this results? Does it mean that these perturbations affect cell survival rates in general instead of affect neural induction specifically?

9. "SERTAD4-AS1 revealed the transgelin (TAGLN) gene - which encodes an actin-binding protein and early marker of smooth muscle cell differentiation". The example presented here is interested, but it would be better to have an example related to neural induction, which is more relevant to the experimental design.

Authors' response to the first round of review

Dear Reviewers,

We thank you and the Editor for reading our manuscript carefully and providing constructive comments. This feedback has been extremely useful in guiding multiple new analyses that have strengthened our conclusions and also for revisions to the text that have clarified our findings. Below, we start with an outline of all revisions and follow this with a point-by-point response for each of the Reviewers' comments.

Sincerely,
Dan Lim

Outline of Revisions:

Updates to Figures:

- Figure 1
  - 1A: Immunocytochemistry staining of iPSCs undergoing neural induction
  - 1B: Flow cytometry analysis of PAX6 protein staining at various stages of differentiation
  - 1C: Heatmap of neural induction transcriptomic expression at various stages of differentiation
- Figure 2
  - 2C: Added labels and color-coding to indicate targets in the individual validation experiment
- Figure 4
  - 4E: New ChIP-Seq signal analysis at promoters of coding and lncRNA genes
  - 4F: New enrichment analysis of coding and lncRNA gene hits in broad H3K4me3 domains associated with genes important in cell identity
- Figure 8
  - 8D: New analysis of *SERTAD4-AS1* knockdown demonstrating no impact on *SERTAD4* coding gene expression (from Perturb-Seq experiment)
- Figure S1
  - S1B: Distribution plot of the □ values of non-targeting control sgRNAs
  - S1C: Analysis of sgRNA agreement for each target, comparing hits vs non-hits
  - S1D: Precision-recall analysis of downsampled data to estimate performance at various levels of screen coverage
- Figure S3
  - S3C: Fixed typos
- Figure S4
  - S4A: New temporal gene expression analysis to illustrate biological hypotheses prior to machine learning
- Figure S5
  - S5A: New analysis of lncRNA gene hit to coding gene hit genomic distances
- Figure S6
  - S6A: New analysis of successful target knockdown in Perturb-Seq experiment
  - S6D-E: Rearranged order to match text with greater cohesion

Response to Reviewers

Updates to the text (colored in blue in the manuscript):

1. Title, abstract, and general changes:

a. Title: Changed from “Ultra-genome-wide screen of coding and lncRNA genome function in human pluripotent stem cells” to “Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells”
b. Removed “multi-omics” from abstract and introduction
c. Added “induced pluripotent stem cells” to the abstract
d. Updated all figure callouts
e. Updated “ultra-genome-wide” to “genome-wide” or “dual genome-wide”

throughout, including on the interactive web resource and GitHub page
f. Converted to Vancouver superscript citation style throughout
g. Simplified language and removed extraneous text whenever possible
2. Results section: Dual genome-wide CRISPRi screens identify coding and noncoding genes regulating neural induction
a. Added text to explain the sgRNA library and design algorithm, including number of sgRNAs per target (lines 108-114)
b. Added text to explain the rationale of the selected time-point (lines 118-122)
c. Added text to explain the calculation of (lines 135-136) and the control distribution (line 140)
d. Added text explaining the sgRNA and screen target coverage (lines 141-142)
e. Added text accompanying new Figures S1C and S1D (lines 146-155)
3. Results section: Validation of screen results
a. Added text to clarify the sgRNAs used in the individual validation (lines 170-172)
4. Results section: Distinct transcriptomics and epigenomics of coding and lncRNA gene hits
a. Added text to better introduce our hypotheses and context before the machine learning analysis (lines 219-226)
b. Added clarification to the definition of individual epigenomic features (lines 233-235)
c. Added clarification to "common transcriptional heuristics" (line 248)
d. Added text to more deeply investigate an example of an epigenomic feature (H3K4me3) and its biological interpretation (lines 251-261)
5. Results section: A small fraction of lncRNA hits have evidence of enhancer-like function
a. Added text regarding the analysis of linear genomic distance of lncRNA gene hits and coding gene hits (lines 266-270)
6. Results section: Dual genome-wide screens enable Perturb-Seq experiment to dissect coding and lncRNA phenotypes
a. Added text to explain Perturb-Seq library design and sgRNA selection (lines 287-292)
b. Added text for analysis of Perturb-Seq knockdown data (lines 303-304)
c. Added text to clarify non-neural cell types that can normally appear during neural induction (lines 316-317)
d. Added text to clarify controls used in Perturb-Seq (line 323)
e. Added text to connect our Perturb-Seq findings as validation for findings from Figure 3 (lines 331-333)
7. Results section: Coding gene repression stalls or aborts differentiation, while lncRNA gene repression permits a greater diversity of cell states
a. Added text to clarify the stalled and apoptotic phenotypes (lines 385-388)
b. Added text to connect Perturb-Seq phenotypes as validation for findings from Figure 3 (lines 404-406)
8. Results section: Repression of *SERTAD4-AS1* increases production of TAGLN+ cells
a. Added text to clarify the neural induction phenotype of *SERTAD4-AS1* (lines 439-444)
b. Added text to explain *SERTAD4-AS1* knockdown without impact on *SERTAD4* (lines 440-442)
9. Discussion
a. Added text to clarify conclusions from machine learning analysis (lines 461-462)
b. Added text to clarify the suspected biological function of *SERTAD4-AS1* (lines 487-488)

c. Added text to substantially discuss the limitations of our study with regards to the use of iPSCs and CRISPRi, and the future steps with regards to genome engineering studies and validation in other pluripotent stem cells (lines 496-507)
10. Methods
a. Formatted according to STAR Methods
b. Added Key Resources Table
c. Added additional methods details to new analyses added in this revision

Point-by-point response:
Reviewer #1:
In the manuscript the authors describe a CRISPRi screen targeting both protein-coding and lncRNA genes, which they term an "ultra-genome-wide" screen. The screen is used to identify lncRNAs with functional roles during differentiation of neuronal cells from embryonic stem cells. The screen provides a nice and important resource of lncRNAs whose perturbation affects the differentiation process and it is followed by follow-up analysis by Perturb-seq and some further description of some of the hits. There are also some general insights about the relative importance of lncRNAs during the differentiation process and reduced importance in proliferating cells. Overall, the paper is of interest to both the lncRNA community and the community interested in CRISPR screens, or more generally, high-throughput approaches for studying lncRNA function. The provided resource should be very useful for the community and the paper sets up high standards for further screen-based studies. As such it should be of interest to the Cell Genomics readership. However, some points need to be addressed before I can recommend publication.
We thank the Reviewer for the assessment of our work and recognizing its value for the community.
Major comments:
1. Since the main focus of the paper is the new CRISPRi library and the screen, several key technical details should be mentioned/discussed in the Results session: How were the promoters of lncRNAs targeted by CRISPRi selected? (how many per lncRNA? How confident are the coordinates?).
The lncRNA targets were selected based on RNA-Seq expression in iPSCs and during neural induction. The sgRNAs targeting these promoters were selected by combining CRiNCL sublibraries (Liu et al., 2017) to cover as many unique targets as possible, with 10 sgRNAs per promoter. Briefly, the design of the CRiNCL sublibrary sgRNAs (discussed in Liu et al., 2017) was based on the revised hCRISPRi gRNA design algorithm version 2.0 (Horlbeck et al., 2016). This algorithm uses FANTOM cap analysis of gene expression (CAGE) data to provide highly confident transcription start site coordinates. These details and citations are now also included in the STAR Methods.
How many of the gRNAs were eventually represented enough in the sequencing data to give reliable depletion/enrichment ratios?
Given the very high screen coverage maintained throughout the screen, >99% of sgRNAs were sufficiently represented in the sequencing data for reliable analysis at the standard threshold of 100X coverage per sgRNA used in prior screens (Gilbert et al., 2014; Horlbeck et al., 2016; Liu et al., 2017). Using a highly stringent threshold (500X per sgRNA), >97% of sgRNAs were represented. This is now stated in the text of the Results section.
How many promoters were covered by those? How many lncRNA genes?
With multiple sgRNAs targeting each promoter, every coding (18,905) and lncRNA (10,678) gene target was covered by multiple sgRNAs passing the threshold mentioned above. 94% of targets were covered by all designed sgRNAs (5 per coding gene and 10 per lncRNA gene) and 99% were covered by at least 80% of designed sgRNAs. This is now stated in the text.
Can this be used to estimate the minimum number of cells that are required for the screen using

the "ultra" library? The authors start from a very large number of cells, it is strictly necessarily retrospectively?

In published studies, the cellular sgRNA coverage typically ranges from 200X to 1000X, with higher coverage providing higher sensitivity at diminishing returns. In our study, we conducted the screen at >1000X coverage. In a new retrospective analysis (Figure S1D, shown below), we randomly downsampled the raw screen data to 10%, 20%, and 50% to approximate different levels of coverage, evaluating performance using precision-recall analysis. At ~100X coverage, precision-recall was poor (AUC 0.52), with only 39% of hits recovered. Performance improved significantly at ~200X (AUC 0.78) and ~500X coverage (AUC 0.86), where more than 80% of hits were identified. We would not recommend 100X coverage, but 200X and 500X coverage would trade-off some sensitivity for significantly increased feasibility.

Did different gRNAs targeting the same promoter tend to correlate with each other (i.e., if one was depleted was the other depleted as well?).

Yes. A concordance score can be calculated for every target in the screen, representing the fraction of correlated sgRNAs for that target (i.e., for a depleted hit, the fraction of independent gRNAs that are depleted). Of hits, 90% had a sgRNA concordance of 80% or greater. This was significantly (p < 1 x 10^{-15}) higher than the concordance of non-hits, which around 50% (indicating no correlated direction). This panel has been added to Figure S1C, with accompanying text in the manuscript.

The authors should also provide a BED file (or better, UCSC browser hub) with the coordinates of the lncRNA they use and those of gRNA target regions.

We have provided a searchable database of the lncRNA annotation with direct links to UCSC Genome Browser on our web portal. We have also uploaded a BED file to GitHub as suggested.

2. It is not clear what the authors define as "dynamically expressed over the time course" (Fig. 1c) and how well the lncRNAs targeted in the screen, are expressed (e.g., in TPM terms) during the time course.

We added a heatmap of gene expression to Figure 1C (reproduced below, left) showing gene expression dynamics over time. We identified dynamically expressed genes using the maSigPro statistical package designed for time-series gene expression analysis, which clusters genes transcripts into distinct temporal patterns (Figure S4C, reproduced below, right). Since the analysis of transcriptomic data is not comprehensively discussed until Figure 4, we have edited the text for cohesion as the number of dynamically expressed genes is not central to the results of Figure 1.

The distribution of expression levels of both coding and lncRNA genes span a wide range (Figure S4B, reproduced below). The expression of lncRNA genes is typically much lower than the expression coding genes, which is consistent with the literature.

Moreover, how often are the genes that appear as targets in the screen are not expressed at all or appear very low during the differentiation time course? E.g., LH09400 which is mentioned as a hit, appears to be expressed at a maximal TPM of 0.11.

Of all targeted genes, 717 coding genes and 739 lncRNA genes did not meet the threshold of detection by RNA-Seq. Of these, 12 were hits (2 coding genes and 10 lncRNA genes, which have not been used as examples in the manuscript).

Can the authors show its RNA-seq coverage? Detect it by some orthogonal methods?

The RNA-Seq coverage of this target is quite low in our dataset, as suggested by the TPM. Using junction-spanning primers, however, we are able to detect the multi-exonic *LH09400* transcript by RT-qPCR at Ct values of 34-35 (red curves, below). A no-RT control of the same samples and primers did not show any signal (green curves). The amplification curves are shown below.

Otherwise it seems that the interpretation that many screen hits are just enhancers seems more likely.

Many studies have found lncRNAs to exhibit far lower expression than protein-coding genes

(Derrien et al., 2012; Djebali et al., 2012). Additionally, standard RNA-Seq methods – even when sequenced to high depth – lack sensitivity in capturing some lncRNAs. For example, RNA CaptureSeq shows far greater sensitivity and is able to detect many lncRNAs not detected through conventional protocols (Mercer et al., 2012).

We recognize that a subset of targets – whether they be lncRNA genes or protein-coding genes – may contain enhancer elements. Since expression level alone does not determine enhancer function, we used multiple lines of evidence to identify hits with possible enhancer-like activity (Figure 5), including the generation of 3D chromatin interaction data as well as analyses of highquality
enhancer datasets such as the extensive FANTOM5 enhancer atlas (Andersson et al., 2014) and a massively parallel reporter assay for enhancers in neural induction (Inoue et al., 2019). We identified a subset of hits (18%) that map to enhancers in these analyses, though *LH09400* was not one of them.

Despite extremely low abundance, lncRNAs can still exert transcript-dependent function. The lncRNA *HOTTIP* (expressed in less than 1 copy per cell) physically interacts with WDR5 to regulate expression of its target genes (Wang et al., 2011). The lncRNA *VELUCT* is a highly unstable transcript that regulates cell viability. Although it is undetectable by standard whole-cell RNA-Seq, it is detectable in the chromatin-bound fraction (Seiler et al., 2017).

3. Is there a tendency for lncRNA screen hits to appear to proximity to coding gene hits? This can also hint at cis-regulatory functions.

Globally, the median distance of lncRNA gene hits to coding gene hits is closer (median 1.4 Mb) than the background distribution of any lncRNA gene to coding gene hit (median 2.0 Mb). This difference met statistical significance (*p* = 0.04), though the distributions are highly overlapping. We have added this as Figure S5A (reproduced below). However, as noted in the response above, multiple other independent lines of evidence were also assessed for *cis*-regulatory functions.

Minor comments:

1. I don't think the "ultra-genome-wide" is a suitable term for the screen conducted. It is true that this screen is somewhat more comprehensive than screens targeting just protein-coding or just lncRNA genes, but since screens targeting either have been performed, and in some cells, both protein-coding and lncRNA gene screens have been conducted, and there are still many transcribed regions, such as eRNAs that are not targeted in the screen. A "dual-codingnoncoding"
screen or something along these lines would be a more suitable name.

We agree with the Reviewer’s suggestion. Incorporating ideas from a few Reviewers, a possible title may be, “Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells.” We are open to feedback on refining the title.

2. Some figures are missing - for example Figure 1D, 2D (in general there discrepancies between the figure numbers and the text).

We sincerely apologize for mislabeling figure callouts and have corrected this in the revision.

3. It is not clear what the authors refer to as an "multi-omics" analysis, since only perturbations followed by sorting or RNA expression are used.

We have removed the term “multi-omics” as it was used imprecisely and was not a major aspect of the study. We had intended to indicate that this work encompassed analyses of a variety of omics data, including CRISPR screening, gene expression (bulk and single-cell), epigenomics (histone marks), and 3D chromatin interactions (PLAC-Seq).

4. The "screen score" used in Figure S1 should be clearly defined in the legend.

We thank the Reviewer for pointing this out and have now clearly defined it in the legend, as well as in the STAR Methods.

5. The rationale for doing the analysis at day 8 should be explained.

We have now clarified the text to explain the rationale and included in Figure 1 the flow cytometry results of PAX6 during different days of neural induction (reproduced below). Our goal was to select a time-point where both positive and negative regulators could be discovered. An earlier time-point may not have allowed time for sufficient differentiation to occur; in such a scenario, sgRNAs that substantially enhanced neural induction would have been enriched, but sgRNAs that prevented neural induction may not have been sufficiently depleted. On the other hand, a later time-point may have precluded the identification of sgRNAs that promoted neural induction and only identified depleted sgRNAs.

6. The 32 gRNAs used in validations are targeting how many promoters/genes? 16 genes? Can the 2 gRNAs targeting the same gene be marked in Fig. 2C? That would be more informative than showing the two replicates.

We thank the Reviewer for the helpful feedback, and have marked gRNAs targeting the same gene with labels and color-coding. Figure 2C has been updated (reproduced below).

7. Figure 5 can be a supplementary figure, as it does not contain any substantial positive results.

We agree with the Reviewer's overall assessment of these findings. We have kept this as a main figure for now, as some of the Reviewer's comments were about *cis*-regulatory function. However, we are open to moving this to a supplementary figure.

8. The authors should elaborate on the criteria used to select the gRNAs for Perturb-seq, and in particular explain what are the "dual hits".

We have now added text in both the main body and the STAR Methods. In brief, we selected the highest scoring hits, excluding those with very high proliferation effects (□ > 1 or < -1) as they would become highly overrepresented or underrepresented in the dataset. Dual hits were any differentiation hits that also produced an impact on proliferation (analyzed in Figure 3), *e.g.*, *POU5F1/OCT4* (discussed in the text for roles in both stem cell divisions and pluripotency).

Reviewer #2:

In this manuscript, Wu et al. have performed a large-scale CRISPRi screening targeting both coding and noncoding genes during neural induction from human embryonic stem cells. This led to the identification of hundreds of coding or noncoding genes that are associated with neural differentiation, which were further interrogated by Perturb-seq. The authors have clearly provided a large amount of data for understanding the characteristics of coding and noncoding genes in cellular function. However, there are also some concerns that the authors need to clarify.

Major points:

1. For the Perturb-seq experiments, only 2 independent sgRNAs for each gene were designed. Although the effects of various targets on neural differentiation including the BAF complexes seem to be convincing as different components of the complexes showed similar effects as shown in the UMAP.

We thank the Reviewer for the recognition that the BAF complexes are a convincing example in the Perturb-Seq analysis. We provide other examples in Figure 6 (PAF1 complex) and S6 (SALL4 regulation of *POU5F1*, Mediator complex, Polycomb complex, and others) that also validate the approach.

However, other targets including SERTAD4-AS1, which was selected for further experimental validation, require more sgRNAs to exclude the off-target effects and obtain a more solid conclusion. In this regard, the authors may also utilize "lncRNA (i.e., SERTAD4-AS1) promoter elimination" or "polyA termination signal" strategies to further validate the function of SERTAD4-AS1.

Although straightforward in concept, basic characterization of a lncRNA gene can require a series of detailed mechanistic studies, such as the case for *lincRNA-p21* (Dimitrova et al., 2014; Groff et al., 2016; Huarte et al., 2010). The suggested strategies, such as polyA signal insertion, require extensive genome engineering and have formed the basis of entire studies on their own (Engreitz et al., 2016; Winkler et al., 2022).

We therefore believe that these experiments would unnecessarily extend the scope of our work without impacting the main conclusions. Our manuscript is ultimately not focused on any one target. Instead, the value of our study is the large resource for the community that highlights broad, unique findings regarding coding and noncoding genome function during differentiation. However, we recognize the value of the Reviewer's suggestions. We now discuss the limitations of our characterization of *SERTAD4-AS1* and the kinds of future studies that will be important. Additionally, we have added a new analysis of *SERTAD4-AS1* knockdown from the Perturb-Seq experiment, demonstrating that while *SERTAD4-AS1* is effectively repressed, *SERTAD4* coding gene expression was not affected (Figure 8D, below, n.d. = not detected; n.s. = non-significant).

2. The authors applied iPSCs rather than human ESCs for the neuronal induction for both CRISPRi screening and Perturb-seq. This raises a concern that the epigenetic memory of iPSCs may predispose iPSCs to specific differentiation related to the parental cell type used for the pluripotency induction. The conclusion drawn by neural induction from iPSCs (especially for some epigenetic regulators related to epigenetic memory in such context) may not be generic for other human pluripotent stem cells. This should be validated using other human pluripotent stem cells or at least be discussed in the discussion part.

We thank the Reviewer for pointing out this important difference between iPSCs and ESCs. In our initial manuscript, we did not adequately discuss these differences. As suggested by the Reviewer and also the Editor, we now discuss the limitations of the iPSC model in the text.

3. Based on the observation that epigenomic data can predict both coding and noncoding hits with similar performance. The authors hypothesized that these lncRNA hits are potential enhancer RNAs. However, further analyses showed that most of these lncRNAs are not enhancer RNAs. What features are these lncRNA hits and what transcripts (lincRNA or others) are these? The authors raised the question without providing any clue in the main text.

We agree with the Reviewer that this is an intriguing finding that was not explored. We have added new text and analyses that investigate these findings more deeply. Specifically, we have provided new analysis on the H3K4me3 histone mark (Figure 4E-F), which was significantly elevated at both coding and lncRNA gene hits compared to non-hits. Further, H3K4me3 broad domains are associated with genes important in cellular identity and function (Benayoun et al., 2014), and we find that such domains are significantly enriched for neural induction screen hits.

4. In Fig 2C, what are replicate 1 and 2? Are they 2 independent sgRNAs for one specific target? If so, authors should include the information that the 2 sgRNA replicates for one specific gene showed similar effects in the screening?

Each target has 2 independent sgRNAs in 2 biological replicates. We have added labels and improved color-coding to the plot to indicate sgRNAs targeting the same gene as suggested. An updated Figure 2C is reproduced below:

5. One advantage of Perturb-seq using CRISPRi is that you can detect both the sgRNA and the expression level of its target in the scRNA sequencing. The authors should show some information in the Perturb-seq whether specific target is efficiently silenced when the sgRNAs are simultaneously expressed, especially for SERTAD4-AS1. This will strengthen your conclusion from the perspective of the screening reliability.

We agree with the Reviewer that this is in theory a key advantage of Perturb-Seq. In our experiment, knockdown was overall robust, showing 80% median knockdown (Figure S6A). Only a small subset of targets (13%) achieved less than 50% knockdown, and approximately half of these were non-hits. Specifically, *SERTAD4-AS1* showed excellent knockdown without significant impact to the *SERTAD4* coding gene (new analysis in Figure 8D, reproduced below, n.d. = not detected; n.s. = non-significant).

6. I would recommend adding neural induction in the title, as the screening was done in the neural induction and not merely in human pluripotent stem cells.

We accept the Reviewer’s recommendation. Incorporating ideas from a few Reviewers, a possible title may be, “Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells.” We are open to additional thoughts to refine the title.

Minor points:

1. "Figure 1D-E" is missing.
2. "Figure 2D" is mislabeled at last line of the part of "Validation of screen results", should be "Figure 2C".
3. "POU5F1" was written wrong at the Figure S3C.
4. "Figure S1B" is mislabeled at third line of the part of "Distinct transcriptomics and epigenomics of coding and lncRNA gene hits"; "Figure 4G" is wrong at fourth line from last. "Figure S1B" is nonexistent.
5. "Figure 6D" is mislabeled at last line of the second paragraph of the part of "Coding gene repression stalls or aborts differentiation, while lncRNA gene repression permits a greater diversity of cell states." "Figure 6D" is mislabeled at the second line from last of the fourth paragraph, should be "Figure 7D".

We thank the Reviewer and sincerely apologize for these errors. There are now all corrected.

Reviewer #3:

In the manuscript titled with "Ultra-genome-wide screen of coding and lncRNA genome function in human pluripotent stem cells", Wu et al. performed a systematic screening targeting around 30,000 coding/noncoding loci based on CRISPRi and identified hundreds of coding and lncRNA genes functional in human neural induction. Very interestingly, they found lncRNA genes are more prone in regulating differentiation rather than proliferation. Next, they utilized Perturb-Seq to further explore these hundreds of primary hits in neural induction. In general, a serial of extensive screening/analysis in this study provided an informative resource about regulators of neural lineage, and more insights about unique function of lncRNAs in lineage differentiation, which should be a strong candidate to publish in Cell Genomics. However, this manuscript was organized very badly and even unbelievably lacked several pieces of figures (i.e., Figure 1D/E and Figure 2D).

More concerns include:

The major concern is the quantitative readout of screening. Since the Figure 1 failed to match with the manuscript, the reviewer did not understand how the authors calculate the rho value (PAX6 staining) for each sgRNA/cell (Figure 1B-C). Was this information obtained from the screen or validation?

We thank the Reviewer for raising this point. We now provide clear details on the rho calculation, which is a common enrichment metric used in marker-based screens equivalent to a log2 fold-change, although with different names in various publications (Adamson et al., 2016; Liu et al., 2017; Parnas et al., 2015). It is obtained from the screen, but a similar calculation can be obtained from validation experiments. Briefly, each sgRNA’s abundance in the PAX6+ and PAX6- fractions are quantified by sequencing and normalized by read depth to counts per million (CPM). The ratio of abundances is then calculated (CPM in PAX6+ fraction / CPM in PAX6- fraction), and this value is log2-transformed to produce rho. This has been clarified in the text and STAR Methods.

Figure 2D was missing, or should be Figure 2C?

We sincerely apologize for the mislabeling and have corrected this in the revision. This callout was meant to refer to Figure 2C.

Figure 3 was of interest. A couple of experimental validation could make the finding more obvious and solid.

We thank the Reviewer for expressing their interest in Figure 3, which illustrates differences in proliferation vs. differentiation effects in the genome-wide screen. Perturb-Seq is independent experiment of hundreds of targets from the genome-wide screen that captures both proliferation and differentiation effects. The higher-resolution findings from Perturb-Seq validated the results of the genome-wide screen. For example, as shown in Figure S6B (reproduced below), Perturb-Seq cell counts (reflecting cell proliferation) were highly correlated to the genome-wide screen growth phenotype □ (Pearson r = 0.92). Additionally, we find coding genes hits to be overrepresented in phenotypes of activating apoptotic programs (Figure 7E, highlighted in red, reproduced below), while lncRNA gene hits were not associated with this pathway. Therefore, the Perturb-Seq experiment provides independent results that support the findings of Figure 3. We have now added text accompanying the results of Figures 6 and 7 to explicitly discuss these validation analyses.

Figure 8 was not that solid and enough to claim that SERTAD4-AS1 functions in neural differentiation. CRISPRi-based knockdown is difficult to distinct the effect on SEARTAD4 and SERTAD4-AS1.

Full dissection of the *SERTAD4-AS1* locus would require genome-engineering strategies such as those mentioned by Reviewer 2 (*e.g.*, promoter deletion and/or polyA insertion). However, given the extensive nature of these approaches that form the basis of studies in their own right (Engreitz et al., 2016; Winkler et al., 2022), we believe that these experiments would unnecessarily extend the scope of our work without impacting the main conclusions, which are broad findings regarding how the coding and lncRNA genomes contribute to neural induction. Based on this comment, we now discuss the limitations of our characterization of *SERTAD4-AS1* in the Discussion. Additionally, we have included a new analysis of *SERTAD4-AS1* knockdown from the Perturb-Seq experiment, demonstrating that while *SERTAD4-AS1* is effectively repressed, *SERTAD4* coding gene expression was not affected (Figure 8D, reproduced below, n.d. = not detected; n.s. = non-significant).

Reviewer #4:

Wu et al developed a novel genome wide screen to determine the coding and noncoding factors that regulate neural induction of iPSC using a large scale CRISPRi technique. The data obtained from this study are novel and informative. It will benefit to our understanding on neural differentiation processes. However, the analysis and validation steps need to be improved substantially. I have following comments:

1. First of all, it is unclear that how rho of 4,523 sgRNAs of non-targeting controls is distributed between PAX6- and PAX6+ population. This is a key information to evaluate the effects of targeting sgRNAs. Because a normal neural induction of iPSC will drive cells to much more toward PAX6+ cells than PAX6- cells, it is not clear if the assumption of "Differentiation phenotypes (rho) were calculated by taking the log2 enrichment ratio of each sgRNA in the PAX6+ versus PAX6- sorted fractions, providing a symmetric measure of the impact on neural induction (as read out by PAX6 protein) on a log-scale" is reasonable or not. This is important as it will affect all downstream analysis and interpretations.

We thank the Reviewer for bringing to attention this important analysis. The control sgRNA rho distribution is symmetric. We have added this analysis to Figure S1B, reproduced below.

2. Although it is interesting to see the PAX6+ population is altered with certain sgRNAs that target designed coding and noncoding targets, it would be more informative to also analyze the top off-target effects of these sgRNAs to ensure the effects of these sgRNAs were from "authentic" target genes instead of "off-target" genes. The GO analysis results (Suppl 3D) indicates that the specificity of the genes identified is not very informative.

Off-target analysis is important for newly designed sgRNAs. The sgRNAs used in this study have been tested, revised, and validated with high off-target activity sgRNAs excluded (Horlbeck et al., 2016; Liu et al., 2017). Nevertheless, we identified potential off-target genes using Cas-OFFinder (Bae et al., 2014) and found that GO analysis of off-target genes did not show any enriched terms. By contrast, the GO analysis of differentiation (Figure S2C) and proliferation hits (Figure S3D) showed many enriched processes specific to each phenotype, as expected. These results indicate that off-target genes are not enriched for any biological processes and that the effects of the sgRNAs are far more likely due to authentic target genes.

3. The experimental validation of the screen is somewhat circling in logic because the same procedure as screening protocol was used. It might be more informative to use several NPC markers other than PAX6 for the validation step.

We provide the validation experiment in Figure 2 as it is a common and expected validation experiment for CRISPR-based screens (Liu et al., 2017; Parnas et al., 2015; Shifrut et al., 2018). We agree with the Reviewer that several markers would be more informative than PAX6 alone. The Perturb-Seq experiment inherently uses many markers and serves as an independent experiment of hundreds of targets that validates the findings from the screen. For example, cells containing positive hit sgRNAs were found to be significantly enriched for differentiated NSC states that expressed many neural markers, e.g., *PAX6*, *FOXG1*, *FEZF2*, and *EMX2* (modified versions of Figures 6D and 7D below). Highly expressed genes from these cells were significantly enriched for gene ontology terms related to neural development, such as forebrain development and neuron fate commitment, indicating that these sgRNAs promoted not just *PAX6*, but an entire neural gene signature (indicated in heatmap).

4. The biological hypothesis and assumption of "Distinct transcriptomics and epigenomics of coding and lncRNA gene hits" are unclear. Did the authors assume the hit genes will express higher in the late stage of neural induction? What does " individual transcriptomic features" mean? i.e. fold changes? Please formulate these features clearly.

We have now clarified all of these points in the main text. Based on prior work (Liu et al., 2017) and common heuristics, our *a priori* hypothesis was that differential expression would be predictive of hits. For example, we hypothesized negative hits would have the expression pattern of PAX6 (high in NSCs, low in stem cells) and positive hits would have the expression pattern of *POU5F1/OCT4* (high in stem cells, low in NSCs). However, when systematically analyzed (Figures 4 and S4), we found that this was not strongly predictive of hits. For example, *PAF1* and *SMARCE1* were strong hits in the screen, yet their usual expression levels during neural induction were stable. This indicated to us the importance of functional screens rather than relying on differential gene expression to infer biological function.

Did the hit lncRNAs will near the individual epigenomic features? What does "individual epigenomic features" mean? i.e. H3K4me3 sites at different differentiation stages? Please formulate these features clearly. What are the biological interpretation of significant AUC values?

We have now explained these points in greater detail and clarity. An example of an epigenomic feature is exactly what the Reviewer suggested: the H3K4me3 level at a specific differentiation stage (e.g., in stem cells). This mark produced significant AUC values for coding and lncRNA genes, indicating that it was enriched at the promoters of hits compared to non-hits (new analysis in Figure 4E-F, below). Non-significant features (e.g., gene expression fold-change at day 8) were those that did not predict hit status over random chance. We have now analyzed the H3K4me3 finding more deeply by assessing H3K4me3 broad domains, which are enriched for hits. We have also included text on its biological interpretation, as broad H3K4me3 domains are associated with genes involved in cellular identity (Benayoun et al., 2014).

5. "The Perturb-Seq library consisted of 480 sgRNAs for 240 unique targets". Please explain how did these targets selected? why were only these sgRNAs were selected among all hits?

We have added clarifying text to this section to address these comments. In brief, we selected the highest scoring hits, excluding those with very high proliferation effects (□ > 1 or < -1) as they would become highly overrepresented or underrepresented in the dataset. We also used random sampling to select non-hits. The total number of sgRNAs were limited by experimental and sequencing cost.

6. "We obtained a total of 78,393 cells that harbored single sgRNA perturbations". Were any cells without sgRNA obtained from single cell RNAseq?

For the Perturb-Seg experiment, we used FACS to obtain as many cells as possible with sgRNAs. Therefore, all cells in this experiment contain sgRNAs. However, 12 unique nontargeting

sgRNAs were included as controls to represent non-perturbed cells, and these formed the largest subset of cells (3,638 cells). Some cells could not be assigned a sgRNA due to dropout during sgRNA capture or sequencing and were excluded (Replogle et al., 2020).

How were the "three major cellular trajectories" determined? i.e. based on the cell population without sgRNA perturbation or all scRNAseq results? if sgRNA perturbation data used, if these data used, will it bias the downstream analysis?

We identified the three trajectories using RNA velocity algorithms based on all cells except for clusters mainly containing only perturbed cells (grey clusters in Figure 6B, reproduced below, left panel). Additionally, analysis of only non-perturbed cells (non-targeting controls) produced the same trajectories (new analysis in right panel, below). Since non-perturbed cells are represented in all trajectories, we do not expect this to bias the downstream analysis.

7. "The group of non-hit sgRNAs were not statistically distinguishable from controls" It is not clear what were "controls" used here?

We used non-targeting sgRNAs as controls. We apologize for not stating this clearly and have added text to the main body and the STAR Methods.

8. "stalled and apoptotic phenotypes (collectively "non-productive") represented the majority (56%) of negative coding hits" What are the explanations for this results? Does it mean that these perturbations affect cell survival rates in general instead of affect neural induction specifically?

The most common phenotype (40%) consisted of cells stalled along the differentiation trajectory. These cells did not have altered survival rates. Instead, they had the gene signature of less-differentiated cells, suggesting that the perturbation impaired in their ability to complete differentiation. Another phenotype (16%) was apoptosis, and these cells likely failed neural induction due to activation of this cellular program. We have now elaborated on these findings in the manuscript.

9. "SERTAD4-AS1 revealed the transgelin (TAGLN) gene - which encodes an actin-binding protein and early marker of smooth muscle cell differentiation". The example presented here is interested, but it would be better to have an example related to neural induction, which is more relevant to the experimental design.

Our original text may have explained the *TAGLN* gene in a confusing way. We have edited the text to better explain the neural induction phenotype of *SERTAD4-AS1*. Briefly, this phenotype occurs in the context of neural induction after repression of *SERTAD4-AS1.* Therefore, the loss of *SERTAD4-AS1* causes an impaired neural induction phenotype. Using Perturb-Seq, we were able to identify that this neural induction phenotype is characterized by the upregulation of *TAGLN*, indicating that *SERTAD4-AS1* may function under normal circumstances to promote proper neural induction by suppressing *TAGLN*.

Referees' report, second round of review

Reviewer #1 (Comments to authors)

The authors have address all of the concerns in a satisfactory way, the manuscript has improved, and so I can now recommend publication.

Reviewer #2 (Comments to authors)

The authors have resolved all my concerns, I am now satisfied with the revisions and recommend acceptance of the manuscript.

Reviewer #3 (Comments to authors)

Most of my concerns have been addressed and it is now acceptable in my view.

Reviewer #4 (Comments to authors)

The authors addressed all my concerns. I have no further comments and recommend for publication.

Authors’ response to the second round of review

Dear Reviewers,

We thank you all again for your constructive comments which have improved our manuscript. We are pleased to hear that all four Reviewers have found our revisions satisfactory and are recommending publication without additional changes.