Long-read RNA sequencing identifies non-coding isoform switching as a regulator of cell fate

Victoire Fort[1,2,3], Gabriel Khelifi[1,2,3], Ruicen He[4,5], Valérie Watters[1,2,3], Melanie Pye[4,5], Nobuko Yamanaka[6,7], Marina Gertsenstein[8], Émeline I. J. Lelong[1,2,3], Jérémy Loehr[3,9,10], Victoria Micha[1,2,3], Jean-Philippe Lambert[3,9,10], Yojiro Yamanaka[6,7], Jeffrey Wrana[4,5], Samer M.I. Hussein[1,2,3*]

[1] Department of Molecular biology, medical biochemistry and pathology, Faculty of Medicine, Université Laval, Quebec City, QC, Canada, G1J 5B3

[2] Oncology Research Program, CHU de Québec-Université Laval Research Centre, Quebec City, QC, Canada, G1J 5B3

[3] Université Laval Cancer Research Centre, Quebec City, QC, Canada, G1J 5B3

[4] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, M5S 3K3

[5] Centre for Systems Biology, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada, M5G 1X5

[6] Rosalind and Morris Goodman Cancer Institute, McGill University; Montreal, QC, Canada, H3A 1A3

[7] Department of Human Genetics, McGill University, Montreal, QC, Canada, H3A 0C7

[8] The Centre for Phenogenomics, Toronto, ON, Canada, M5T 3H7

[9] Department of molecular medicine, Faculty of Medicine, Université Laval, Quebec City, QC, Canada, G1J 5B3

[10] Endocrinology and nephrology Research Program, CHU de Québec-Université Laval Research Centre, Quebec City, QC, Canada, G1J 5B3

* Correspondence to: Samer M. Hussein, Ph.D.
Department of Molecular biology, medical biochemistry, and pathology

Université Laval Cancer Research Centre

Oncology Research Program - CHU de Québec-Université Laval Research Centre

Hôpital de l'Enfant-Jésus - Pavillon L

1401, 18e Rue

Québec (Québec) G1J 1Z4

Telephone: (418) 525-4444 ext. 67945

Email: samer.hussein@fmed.ulaval.ca

Abstract

Development of long-read RNA sequencing technologies has paved the way to the exploration of RNA isoform diversity and its relevance in regulating cell fate. However, identifying new functional isoforms is still very difficult. Here, we leverage long-read RNA sequencing to study changes in isoforms during somatic cell reprogramming and identify novel isoforms occurring throughout cell state transitions. We demonstrate tight regulation of non-coding isoforms and show that isoform switching plays previously overlooked functional roles and outcomes in gene regulation and cell fate changes. We uncover a novel long non-coding RNA, *Snhg26*, that undergoes isoform switching during reprogramming to enhance the conversion of differentiated cells towards the pluripotent state. Knock-down of *Snhg26* in mouse and human pluripotency models reveals that it is important for pluripotency acquisition. Together, our study provides a resource to study full-length isoform usage during cell fate change, and demonstrates the power of long-read sequencing to identify functionally relevant gene isoforms in the context of cell plasticity.

Introduction

Among high-throughput technologies, the profiling of RNA transcripts, known as transcriptomics, has become a gold standard to identify the molecular markers that define cell identities and plasticity[1,2]. Despite the extensive collection of RNA sequencing (RNAseq) datasets generated by various independent research teams and international consortiums[3-6], technical limitations have hindered our efforts to obtain a comprehensive and detailed overview of full transcriptomes. Notably, long coding and non-coding transcripts expression, repetitive sequences and transposable elements expression, as well as isoform diversity have been largely overlooked. Indeed, short-read RNAseq cannot cover the full-length sequence of a transcript and thus doesn't have the power to unravel complex locus sequences and isoform diversity[7,8]. Long-read RNA sequencing (lrRNAseq) technologies, that are able to identify full-length transcripts[9,10], can overcome this issue and deepen our understanding of cell biology and organism development. For instance, Leung *et al.* have uncovered the existence of novel isoforms in the context of

both human and mouse cortices, and highlighted changes in alternative splicing between human foetal and adult cortices[11]. On the other hand, studies in human[12] and mouse[13] pre-implantation embryos have described thousands of unannotated isoforms, transcribed from both known and novel genes. In humans, Torre *et al.* have shown significant modulations of alternative splicing events associated with disruption of open reading frames during embryonic genome activation[12]. In the mouse, Qiao *et al*. have leveraged their lrRNAseq data analysis to identify new transcript isoforms where their depletion negatively impacted embryo development[13]. Surprisingly, aside from the latter study, none of these generated datasets have been used yet to identify functionally relevant new gene isoforms in a biological context.

Among the turning points of embryonic development, the establishment of pluripotent stem cells (PSCs) and their differentiation towards other cell types is crucial. Understanding the regulatory network of PSCs has enabled the seminal work by Yamanaka *et al.*, which has led to the discovery that a small set of genes (*Oct4, Klf4, c-Myc* and *Sox2*, termed "OKMS") can reprogram cells into induced pluripotent stem cells (iPSCs)[14]. Since then, extensive studies have been conducted to understand the principles driving this process of cellular plasticity by characterizing reprogramming and PSCs in terms of epigenetics[15,16], transcriptomics[16-18], proteomics[19,20], transcription factor networks[21,22] and metabolomics[23,24]. This body of work has established that induction and maintenance of the pluripotent state are regulated by a gene regulatory network composed of hierarchical and interconnected regulatory circuits mediated by extrinsic signals, key transcription factors, chromatin modifying enzymes, microRNAs and long non-coding RNAs (lncRNAs)[25-27]. Additionally, post-transcriptional regulation (e.g. RNA splicing) has been identified as a potential player in the regulation of reprogramming towards the pluripotent cell state[28-30]. Indeed, short-read RNAseq analyses of mouse reprogramming has demonstrated that splicing patterns from the somatic cells are converted into that of pluripotent cells during iPSC induction, and that RNA-binding proteins (RBPs) regulate alternative splicing in a temporal manner[31,32]. Additionally, overexpressing or knocking-down splicing regulators enhanced or suppressed the efficiency of the reprogramming process, validating their importance in pluripotency regulation[31,33-35]. Despite the

admitted role of alternative splicing and splicing regulators in reprogramming, isoform diversification that inherently comes with alternative splicing events has never been investigated in this context.

In this work, we characterize in detail stem cells as well as for the first time the process of reprogramming mouse somatic cells to iPSCs using lrRNAseq. We identify multiple new isoforms that result from previously undetected splicing patterns, and that are differentially regulated during the reprogramming process. More specifically, non-coding isoforms undergo negative regulation during the mid-phases of reprogramming, and about 6-16% of spliced genes exhibit a switch in expression of their isoforms. Among these, long non-coding RNAs are underrepresented, which suggests lncRNAs undergoing isoform switch may have important functions during the acquisition of pluripotency. We demonstrate that overexpression of a specific isoform of a lncRNA, *Snhg26.217*, enhances the reprogramming to the pluripotent state. We further show that *Snhg26* knock-out results in defects of an *in vitro* blastoid model, but *Snhg26.217* rescues the observed developmental phenotype. Conversely, we show that knock-down of this lncRNA in mouse embryonic stem cells (ESCs) primes cells for differentiation, validating its function in the regulation of pluripotent stem cells. Finally, we show that knock-down of the human ortholog of *Snhg26* leads to a similar outcome in human pluripotent stem cells and also decreases the efficiency of human reprogramming. These results highlight the discovery of a new conserved lncRNA involved in the regulation of the pluripotent state. Together, our study demonstrates the benefits of lrRNAseq to identify new functionally relevant gene isoforms, in the face of exploring isoform diversity. The dataset we generated also stands as a valuable resource for other researchers to study the role of alternative splicing and isoform diversity in the regulation of cell fate.

Results

Defining reprogramming stages by lrRNAseq

Addressing isoform diversity in reprogramming is hindered by the low efficiency of the process[36]. Here, we designed two reprogramming systems that allow generation of sufficient amounts of

reprogramming cells for high resolution analysis of isoform diversity. Because overactivation of *c-Myc* induces transcriptional and splicing events caused by its effect on promoting proliferation[37-39], we devised our systems where *c-Myc* can be either included (OKMS) or omitted (OKS) from the cocktail of reprogramming factor transgenes. This way, reprogramming-induced events that are independent of *c-Myc* can be isolated (please refer to **Supplementary Note** and **Supplementary Fig. 1** for more details on the system). Briefly, we generated stable cell lines of mouse neural progenitor cells (NPCs) with doxycycline (DOX)-inducible reprogramming factor transgenes driving the expression of an *mCherry* reporter gene (**Fig. 1a** and **Supplementary Fig. 1a-b**). As transgene silencing is a necessary step to reach *bona fide* pluripotency at the end of reprogramming[40-42], DOX is removed at D6, and cells able to fully reprogram convert to iPSCs (please refer to **Supplementary Note** for more details on DOX removal). To monitor the appearance of iPSCs, the NPC lines also harbor a *GFP* transgene under the control of endogenous *Oct4* gene promoter expression, which is activated when cells reach the iPSC state (**Fig. 1a** and **Supplementary Fig. 1a-b**). In parallel, we also maintain cells with DOX during the whole reprogramming process to reproduce other previously shown reprogramming paths[16]. For these cells, transgene silencing naturally occurs in late timepoints[43], causing a decrease in mCherry-positive cells (**Supplementary Fig. 1a-b**). Therefore, we collected samples with DOX at D2, D4, D6, D10, D14, D18, and D20, We also collected iPSC lines after DOX removal at either D14, D18, or D20 (**Fig. 1a**).

To validate our reprogramming systems, we performed whole transcriptome analysis using lrRNAseq at different timepoints (n=45 total samples) of the process (**Fig. 1a** and **Supplementary Table 1**). Analyzing expression of reprogramming and PSC marker gene sets (previously described[16]), OKMS reprogramming rapidly activated immediate responder genes, intermediate reprogramming genes, genes associated with mesenchymal-to-epithelial transition (MET) (**Fig. 1b** and **Supplementary Table 1**), all of which are signs of proper of reprogramming[16]. Early and late ESC-like genes were activated as early as D4 and D6, respectively (**Fig. 1b**). This was accompanied by decreased expression of epithelial-to-mesenchymal (EMT) genes starting at D4 (**Fig. 1b**) suggesting our system recapitulates MET, a hallmark process observed during reprogramming to iPSCs[44,45]. OKS reprogramming showed slower dynamics,

and the later expression of the gene sets compared to OKMS reflects the well-known delay in OKS kinetics and is due to lack of *c-Myc*, which is known to enhance reprogramming by increasing proliferation[46,47]. Nonetheless, OKS reprogramming showed rapid increase in expression of immediate responder and intermediate genes, and to some extent MET genes within the first 4 days (**Fig. 1b**) when compared to NPCs. These results suggest that both our OKMS and OKS systems constitute efficient reprogramming strategies, which faithfully recapitulate the global gene expression changes observed in other studies.

Because of the differences in OKMS and OKS kinetics, we undertook hierarchical clustering analysis in order to group OKMS and OKS timepoints that are similar based on expression of the top 10% varying genes (**Fig. 1c**). This segregated the samples into 5 different categories: First the "NPC group" that gathers four replicates of wild-type NPCs, the OKMS-NPCs and the OKS-NPCs at D0. OKMS and OKS reprogramming cells at D2 were classified as an "early reprogramming group", while D4 to D6 for OKMS or D4 to D10 for OKS clustered together and form the groups of "middle reprogramming" phases. OKMS cells between D10 and D18, and OKS cells at D18 and D20 are clustered in a separate "late reprogramming group" that resembles a "PSC group", which is composed of ESCs, OKMS- and OKS-iPSCs in 2i-LIF media. This is consistent with the OKMS samples being quickly redirected towards the induced pluripotent state and achieving pluripotency reacquisition earlier than OKS samples (**Fig. 1b** and **Supplementary Note 1c**). Finally, a sixth group comprising samples of NPCs placed in FBS-LIF medium without reprogramming factors ("FBSL") (**Fig. 1c**) was used to assess transcriptional and splicing changes induced simply by the change of culture media. The segregation of these groups based on reprogramming phases was used in the subsequent analyses of our dataset. Looking at global gene expression, we observed a two-wave pattern (at the beginning of reprogramming upon DOX induction, and at the late phase), highlighting again known reprogramming dynamics, as previously described[40,48] (**Supplementary Fig. 1c** and **Supplementary Table 1**).

Finally, we validated that iPSC clones isolated from our OKMS and OKS reprogramming experiments were able to contribute to the formation of chimeric mice when aggregated with 8-cell stage

mouse embryos. The clones contributed 10-50% to chimera formation, as judged by coat color (**Fig. 1d**). This confirmed that our reprogramming systems give rise to *bona fide* iPSCs. Altogether, these results suggest that we generated both *Myc*-dependent and *Myc*-independent reprogramming systems that follow the well-accepted kinetics of reprogramming and can be compared in further steps of analysis.

Unprecedented isoform diversity and tight regulation of non-coding isoforms

Alternative splicing has been extensively characterized in pluripotent stem cells and during reprogramming[28,32,49]. However, a detailed characterization of transcriptomes from the different cell states found on the trajectory of the reprogramming cells has not yet been undertaken. In order to analyze isoform diversity and confidently identify functional isoforms in our reprogramming systems, we combined single transcriptomes built from each of the 45 samples that we sequenced. We identified 279,998 isoforms that can be reliably detected and supported by several criteria (**Supplementary Fig. 2-5**, **Supplementary Tables 2-3**, and **Supplementary Data 1-3**). Among these criteria, we focused on isoforms that are sufficiently expressed (≥ 1 transcripts per million reads (TPM) in at least 3 samples), have a poly(A) motif, and have no reverse transcriptase switching events (**Supplementary Fig. 2** and **Supplementary Fig. 3a**), where the transcription start site (TSS) was supported by orthogonal approaches (e.g. CAGEseq, 5' end sequencing, H3K4me3 ChIPseq peaks, etc.) (**Supplementary Fig. 3b-c** and **Supplementary Fig. 4**), and where splice junctions were supported by short-read RNAseq, presence of repeat elements, or mouse ESCs proteomics data (**Supplementary Fig. 3d-g** and **Supplementary Tables 2-3**). We also compared our transcriptome to ones built by three orthogonal transcriptome-building tools: Bambu[50], IsoQuant[51], and StringTie[52]. Strikingly, all tools showed ~50% reproducibility when compared to one another (**Supplementary Fig. 3h**, **Supplementary Table 3**, and **Supplementary Data 3**). This is concordant with a recent benchmarking study by the Long-read RNA-Seq Genome Annotation Assessment Project[53]. Lastly, to assess whether non-supported isoforms (7% not supported by TSS (**Supplementary Fig. 3b**), ~20% not supported by junctions (**Supplementary Fig. 3d**)) should be kept or not, we evaluated several metrics of isoform quality (i.e. length, number of exons,

expression and coverage at the lowest-covered splice junction in short-read datasets) and show that non-supported isoforms do not exhibit any major differences from supported isoforms (**Supplementary Fig. 5**). Thus, we also kept those isoforms.

Next, we used SQANTI3 software[54] to classify the 279,998 isoforms. From these, only 35.3% of isoforms were annotated, whether they matched completely (20.51% "Full Splice Matches" or FSM) or partially (14.80% "Incomplete Splice Matches" or ISM) to the reference transcriptome (**Fig. 2a** and **Supplementary Table 2**). Indeed, reprogramming is a process where many cells undergo rapid proliferation, a mesenchymal-to-epithelial transition, DNA damage, global chromatin marks reorganization and a change in cell type, all of which can greatly alter the transcriptome[55]. Thus, this could explain the identification of the remaining 181,127 (64.70%) novel isoforms. From the total, 60.7% of the transcripts were new isoforms from known genes, but with differing splice-junction combinations, with 15.95% coming from new combinations of known splice junctions ("Novel-In-Catalog" or NIC) and 44.75% from novel splice junctions ("Novel-Not-in-Catalog" or NNC) (**Fig. 2a**). The remaining 4% of the detected isoforms were new transcripts generated from novel genes (i.e. 11,169 isoforms), with almost two third of these originating from antisense (1.32%) and genic (1.18%) isoforms, whereas intergenic (0.99%) and fusion (0.50%) transcripts accounted for the rest (**Fig. 2a**). These results are consistent with previous transcriptome analyses using lrRNAseq[11,13,54].

Investigating subcategories of isoforms (**Fig. 2b**), we found that NNC isoforms with at least one novel splice site constituted the main subcategory (41.03%), followed by isoforms with shorter 3' end (FSM *alternative 3' end* and ISM *5' fragment* subcategories, at 7.75% and 7.63%, respectively) and isoforms with *retained introns* (11.97% in total) (**Fig. 2b**). Interestingly, on average 48% of isoforms classified as having intron-retention were predicted to undergo nonsense-mediated decay (NMD) (NIC (49.38%), NNC (45.28%) and fusion (50.67%) *intron retention* subcategories) (**Fig. 2c** and **Supplementary Table 2**). This is consistent with the role of intron retention in NMD[56], but more importantly, it recapitulates what we previously observed as a role for intron retention in regulating gene expression during reprogramming[16]. Taken together, our results highlight the importance of lrRNAseq in

identifying new isoforms and confirm previous findings that the transcriptome undergoes significant splicing and RNA processing events during the reprogramming process, but most importantly, they suggest that isoform diversity has precedingly been underestimated[31,35].

We next analyzed for coding potential and found that the proportion of non-coding isoforms was significantly lower than coding isoforms (15.3% compared to 84.7%, respectively) (**Fig. 2d** and **Supplementary Table 2**), and this is was especially true for known genes (FSM: 86.38%; ISM: 83.10%; NIC: 91.43%, and NNC: 85.88%, with coding probability medians between 0.89 and 0.98). Moreover, non-coding isoforms came predominantly from novel genes, with intergenic and antisense isoforms showing a very low coding score (median of 0.08 and 0.09, respectively, **Fig. 2d**), and almost 80% of these isoforms being non-coding (79.73% and 79.42%, respectively). This is consistent with intergenic regions generally harboring lncRNAs[57,58] and with most antisense genes being transcribed as lncRNAs[59,60]. Given that the majority of non-coding isoforms arise mainly from the relatively small number of novel genes, this suggests that for known genes there is a preference for coding isoforms and a selection against non-coding isoforms.

To determine how the dynamics of non-coding isoform diversity is affected during reprogramming, we segregated isoforms by reprogramming phase, based on the groupings described in **Fig. 1c**. (**Supplementary Table 2**). Overall, we found a significant decrease of the proportion of non-coding isoforms in the middle phase of OKMS reprogramming, which was also reflected in the final iPSC stage (**Fig. 2e**). Conversely, the proportion of coding isoforms increased. These opposing changes were mostly driven by novel isoforms from both known and novel genes (i.e. NIC, NNC, and novel genes) but not annotated genes, with a striking decrease in non-coding NNC isoforms (**Fig. 2f**). This was not due to bias introduced by the use of different types of sequencing platforms or flow cell technologies (i.e. MinION vs PromethION, and R9 vs R10 flow cells, respectively), as the bias was only limited to novel but not known genes (**Supplementary Fig. 6a-b**), and the majority of novel isoforms come from known genes. Altogether, our data expand on previous knowledge by characterizing for the first time both coding and non-coding isoforms in the context of reprogramming. Our results suggest that there is tight and

controlled regulation of isoform diversity during reprogramming with a selection against novel non-coding transcripts.

Alternative splicing follows a biphasic pattern during reprogramming

As NNC isoforms are transcripts that use novel donor and/or acceptor splice sites to create new splicing junctions, we aimed to investigate whether the observed changes in the abundance of non-coding transcripts could be explained by post-transcriptional regulation mechanisms such as alternative splicing (AS). Indeed, AS has been demonstrated as one of the underlying mechanisms affecting reprogramming efficiency, with different RBPs coordinating AS events at each timepoint of the process[32]. Thus, we first assessed expression of 19 splicing regulators known to modulate alternative splicing in pluripotent stem cells and reprogramming[32,35], splitting them into 4 groups based on their expression: Increased expression during reprogramming to pluripotency (group 1), expressed throughout (group 2), decreased (group 3), or transiently increased (group 4) (**Fig. 3a** and **Supplementary Table 4**). The expression profiles of these RBPs did not globally differ when comparing the two reprogramming systems, although it highlighted again delayed kinetics in OKS reprogramming compared to the OKMS system, and was consistent with previous studies[31,32,35,61]. For examples, *Tia1*, *Mbnl1* and *Mbnl2* expression was down-regulated (group 3) in both systems, which agrees with their known function as negative regulators of ESC-like AS patterns[32,61,62]. Similarly, *Esrp1* and *Esrp2* expression transiently peaked between D6 and D10 for OKMS, and between D18 and D20 for OKS (group 4), following the same pattern as observed with markers of the MET (**Fig. 1b**). This is in line with their known role in promoting AS changes at the MET stage of reprogramming[32,63]. Thus, our results recapitulate the expression changes of regulators known to coordinate temporal AS patterns observed during induction to pluripotency.

To further investigate splicing regulation, we assessed AS events (defined in **Fig. 3b**) during reprogramming. For this, we analyzed our 279,998 identified isoforms (**Supplementary Fig. 2** and **Supplementary Table 2**) and removed events occurring when comparing NPCs to FBSL samples, which

represent changes due to media change only, to focus on events only induced by reprogramming factors (**Supplementary Fig. 7a-b**). With this in mind, we detected a total of 164,341 AS and 102,550 differential isoforms usage events occurring throughout the reprogramming process (**Supplementary Fig. 2** and **Supplementary Table 4**). The majority of AS events, with alternative first exons dominating the events, occurred when NPCs transitioned to the early stages of reprogramming (n=13,974 for OKMS; n=12,300 for OKS), irrespective of the reprogramming system (**Fig. 3c**). AS events were less included during the early stages of reprogramming (NPCs to middle stage: A first phase), whereas the inclusion of splicing variations increased steadily in the later stages (**Fig. 3c**) (middle to late/iPSC stages: A second phase), suggesting a biphasic pattern that switches in the middle of reprogramming to include splicing events.

To investigate if these reprogramming specific AS events could lead to functional consequences, we performed an isoform switching (IS) analysis. Isoform switching can inform of AS events that lead to open reading frame (ORF) disruptions or changes that affect protein expression. IS can be measured by calculating differential isoform usage using differential isoform percentage spliced-in (ΔPSI) events ("ΔPSI events-based method") or TPM values to obtain differentially expressed isoforms (DEIs) ("DEIs-based method"), where the number of ΔPSI events or DEIs is proportionally indicative of the number of IS events. Using the ΔPSI events-based method, we found that IS peaked early during reprogramming (NPCs to early/mid-reprogramming), with a higher number of less abundant events (~4,000) than more abundant events (~2,000) (**Fig. 3d**). Similar to AS events, IS was also biphasic, changing from less abundant in the early stages of reprogramming to more abundant in later stages (**Fig. 3d**). Interestingly, less abundant isoforms in early and middle phase reprogramming showed lower coding probabilities (i.e. decreased or disrupted ORFs) than more abundant isoforms (**Fig. 3d**). This is consistent with the specific decrease in non-coding isoforms in the middle phase of reprogramming (**Fig. 2e-g**). In parallel, a global decrease was observed in 3' untranslated region lengths for alternatively spliced isoforms that are up-regulated in those later reprogramming stages (**Fig. 3e**, **Supplementary Fig. 7c-d**, **Supplementary Table 2**, and **Supplementary Data 4**), which has been shown to be associated with higher protein-coding

potential[64]. Taken together, these results suggest that AS leads to gain of protein-coding potential and function, and tight regulation of non-coding transcripts in the early phases of reprogramming.

Selective interplay between gene expression and splicing regulates reprogramming

Having identified major splicing and isoform usage changes during reprogramming, especially in the early phases, we investigated how this affects isoform expression and gene function, and whether common or different genes are affected in other datasets. In total, we identified 10,490 genes that undergo splicing changes during reprogramming, and of these 49 genes (among them *Tpm1, Mbnl2, Zfp207, Fgfr1, Palm, Fn1, Dst, Fyn*) were found in common between our dataset and two other reprogramming datasets[32,35] (**Fig. 4a** and **Supplementary Table 4**). Strikingly, we found that more than 85% of the genes identified as differentially spliced in these datasets overlapped with our dataset, whether the cells used for reprogramming were NPCs (our study), mouse embryonic fibroblasts (MEFs) (637 genes out of 676, i.e. 94.2%) or B cells (131 genes out of 148, i.e. 88.5%). However, less than 7% of the genes undergoing AS we detected were found in the two other studies. Thus, our results highlight the strength of using lrRNAseq to detect novel isoforms and splicing events compared to other technologies.

Of the 49 common genes, we found that *Mbnl2*, a negative regulator of AS in pluripotency[32,61,62] (**Fig. 3a**), undergoes two exon inclusion/skipping events during reprogramming: Inclusion of exon 6 yields a bipartite nuclear localization signal, thus exon 6 skipping can affect the localization and function of MBNL2[65]. Similarly, skipping exon 9 causes a frameshift and decreased protein stability[65](**Fig. 4b** and **Supplementary Table 4**). Interestingly, exons 6 and 9 were specifically skipped in both OKMS and OKS reprogramming from the early phase (**Fig. 4c** and **Supplementary Fig. 8a**), suggesting loss of function of MBNL2. Those exons were then included in OKMS mid-reprogramming, and the global gene expression remained stable (**Fig. 4c-d**). As for OKS, expression strongly decreased in the late phase, suggesting that different mechanisms of gene regulation are selectively used between the two systems (i.e. splicing versus expression changes in OKMS and OKS, respectively). Other commonly spliced genes are involved in cell

signaling and adhesion (**Fig. 4e** and **Supplementary Table 4**). Among them are negative regulators of pluripotency, such as tyrosine kinases *Fgfr1* and *Fyn* (**Fig. 4a** and **Fig. 4e**), which promote differentiation of ESCs[66]. *Fyn*, a SRC family kinase, contains a mutually exclusive exon 7 (exons 7a and 7b), that leads to differences in FYN SH2 linker and N-terminal of the catalytic domain SH1 (**Fig. 4f** and **Supplementary Table 4**). Inclusion of exon 7a decreases FYN accessibility to its ligand and thus results in lower activation of the kinase[67]. We found that both decrease in gene expression and inclusion of exon 7a could lead to a decrease in FYN functionality during reprogramming, more specifically OKS reprogramming (**Fig. 4g-h**, **Supplementary Fig. 8b-c**, and **Supplementary Table 4**). Interestingly, expression of genes negatively-correlated with *Fyn* (based on DepMap[68] analyses) increased in the early and middle phases of reprogramming following the inclusion of exon 7a in OKS (i.e. gain of inactive form of FYN, thus behaving like loss of FYN). Comparatively, no significant changes were observed in OKMS, consistent with the low exon 7a inclusion, but there is a decreasing trend in the expression of those FYN targets (**Fig. 4i**, left and **Supplementary Table 4**). As for positively-correlated genes, we noticed an increase mostly in the middle and late phases of OKS and OKMS reprogramming, respectively, where only the active form of FYN is expressed (i.e. exclusion of exon 7a and inclusion of exon 7b) (**Fig. 4g-h** and right panel of **Fig. 4i**). These results suggested that expression of an inactive isoform of FYN at the beginning of reprogramming corresponds accordingly with the regulation of its target genes.

To investigate possible functional effects of alternative splicing on a global level during reprogramming, we assessed the proportion of isoforms that included PFAM protein domains, as well as non-coding isoforms that were generated from protein-coding genes (**Fig. 4j-k** and **Supplementary Data 4**). We observed an increasing trend in isoforms containing annotated PFAM domains throughout both the OKMS and OKS reprogramming timelines (**Fig. 4j**), which is consistent with the higher coding potential of more abundant isoforms versus less abundant isoforms during reprogramming (**Fig. 3d**). Inversely, the proportion of non-coding isoforms from protein-coding genes decreased significantly during reprogramming (**Fig. 4k**). This pattern was true for all expressed genes and did not depend on

alternative splicing. Taken together, our results are consistent with the notion of selection against non-coding isoforms during reprogramming and suggest that a broader selection of functional isoforms (i.e. isoforms containing PFAM domain) is at play to ensure proper reprogramming to pluripotency. This again highlights the strength of using lrRNAseq to identify isoforms which are functionally relevant for studying specific biological processes.

Isoform switching selectively regulates non-coding isoforms during early reprogramming

Given that both expression and AS selectively regulate reprogramming, we sought to determine if these splicing patterns affected isoform expression. From the 279,998 identified transcripts, we focused only on those that are differentially expressed in at least one sample comparison in our dataset. This refined the list to 198,247 isoforms (**Supplementary Fig. 2** and **Supplementary Table 5**), which are significantly more or less expressed during the different phases of reprogramming (**Fig. 5a**). Similar to our splicing analysis, we excluded isoforms affected by media change only (FBSL to NPCs comparison, **Supplementary Fig. 9a**). Similar to differentially expressed genes (**Supplementary Fig. 1c**), we observed a 2-wave pattern of DEIs, where more isoforms are up-regulated in the early and lates phases of reprogramming (much more pronounced in OKS system) and less DEIs between the early and middle phases of the process (**Fig. 5a**). This held true for both coding and non-coding isoforms (**Supplementary Fig. 9b**). This also suggests that the early phases of reprogramming are in part conserved in terms of isoform expression regulation and coincides with the fact that several molecular and cellular mechanisms important for reaching the pluripotent state take place in the early steps of reprogramming[40,44,69,70]. Therefore, we narrowed our subsequent analyses to isoform expression changes occurring as the cells progress from NPCs to mid-reprogramming. We classified isoform changes into 4 categories based on their expression profiles: Up-regulated ("up"), down-regulated ("down"), up- then down-regulated ("up-down"), or down- then up-regulated ("down-up") (**Supplementary Fig. 9c**). This reduced our list of DEIs to 35,694 transcripts (18% of total DEIs, corresponding to 12,501 genes) (**Supplementary Fig. 2** and

Supplementary Table 5) but also allowed to define and track isoform expression patterns. For example, we found that isoforms that were up-regulated in early to mid-reprogramming were later down-regulated in late reprogramming (n=9,650 for OKMS; n=6,964 for OKS) (**Supplementary Fig. 9c**).

With this refined DEIs list and given that both gene expression and AS are selectively regulated as we previously showed, we now sought to integrate the notion of isoform expression with proportionality in our IS analysis (i.e. "DEIs-based method"). This is different from our ΔPSI event-based IS analysis (**Fig .3d**), which mainly represents isoform proportions. For this we devised an in-house matrix to recapitulate the different possible scenarios of IS occurring between NPCs and mid-reprogramming (**Fig. 5b** and IS matrix in **Supplementary Table 5**). We found 2,081 and 815 genes undergoing IS for OKMS and OKS, respectively, affecting about 16.6% and 6.5% of genes where isoforms are differentially expressed between NPCs and mid-reprogramming (**Supplementary Fig. 9d** and **Supplementary Table 5**). GO-term enrichment analysis showed that genes undergoing IS between NPCs and mid-reprogramming for both OKMS and OKS were associated with processes such as mRNA processing, RNA splicing, apoptosis, cell division, DNA repair, chromatin remodeling, translation, cellular respiration and stem cell population maintenance and differentiation (**Fig. 5c**, **Supplementary Fig. 9d**, and **Supplementary Table 5**). This is expected as these biological mechanisms are key actors in induction of pluripotency. Thus, isoform switching is involved in every layer of pluripotency induction regulation, and this is conserved in our two different reprogramming systems.

Interestingly, while no differences in the patterns of DEIs could be detected between coding and non-coding genes (**Supplementary Fig. 9b**), we observed a striking lack of non-coding genes undergoing IS (**Fig 5d**). Indeed, around 96% of IS genes corresponded to protein-coding genes, whereas only about 2-3% were lncRNA genes, and around 1% constituted pseudogenes or other biotypes. We compared these numbers to the proportion of biotypes in the list of genes undergoing differential isoform expression but not necessarily IS (n=12,501 genes for OKMS; n=8,476 genes for OKS) and found that lncRNAs were significantly less enriched within the IS gene list: lncRNAs were expected to amount for 12-13% of genes, but they represented only 2-3% in our datasets for IS (**Fig 5d**). In contrast, protein-coding genes

were significantly more enriched within the IS gene list (96.6% for OKMS; 96.1% for OKS) compared to expected values (83.3% for OKMS; 84.8% for OKS). These results suggest that there is a negative constraint against lncRNAs to undergo IS, consistent with our results demonstrating a constraint and tight regulation of non-coding transcripts during the early phases of reprogramming (**Fig. 2e-g**).

When looking at genes classified as undergoing IS by both with the ΔPSI events-based and DEIs-based methods of IS analysis (n=1,766 genes) (**Fig. 5e** and **Supplementary Fig. 2**), we observed that, similar to the previous patterns for all genes (**Fig. 4j-k**), reprogramming is associated with an increase in the proportion of isoforms containing annotated protein domains (**Fig. 5f**), and a concomitant and significant decrease in the number of non-coding isoforms. Interestingly, genes subject to IS exhibited higher proportions of both isoforms with annotated protein domains (i.e. putative functional isoforms) and non-coding isoforms across reprogramming samples compared to all expressed genes or genes subject to AS (**Fig. 5g**). This suggests that IS allows for specific selection of both functional coding isoforms, encoding annotated protein domains, and non-coding isoforms from protein-coding genes.

One example of a protein-coding gene with demonstrated non-coding isoforms regulation during reprogramming is *Srsf2*. *Srsf2* is a well-known regulator of AS in early reprogramming[31,71], and was among the genes in the main GO-terms associated with IS (i.e. "RNA splicing" and "mRNA processing") (**Fig. 5c**). *Srsf2* gene-level expression did not significantly change during both OKMS and OKS reprogramming but tended to increase when cells reach the pluripotent state (**Supplementary Fig. 9e**). We identified 13 DEIs for *Srsf2*, six coded for the proper protein, five were predicted to undergo NMD while retaining a strong coding probability, and two were non-coding (**Fig. 5h**). Of the five NMD transcripts (NMDex3), three were NIC (new combinations of known splice junctions) isoforms that included a splicing in exon 3 that was distinct from the protein-coding (main) isoforms (**Supplementary Fig. 5h** and **Supplementary Table 6**). We compared expression profiles and fractions of the main, NMDex3, and non-coding isoforms. In OKMS reprogramming, the proportion of the protein producing isoforms and the NMDex3 isoforms decreased significantly (going from 58.6% and 38.1% in NPCs to 32.9% and 5.2% in middle phase, respectively) (**Fig. 5i** and **Supplementary Fig. 9f-g**), when their

expression levels were not significantly altered. This was due to a significant increase in the expression and thus proportion of the non-coding isoforms (going from 3.3% in NPCs to 61.9% in middle phase) (**Fig. 5i**, **Supplementary Fig. 9f-g**, and **Supplementary Table 6**). This pattern was reversed in the late and iPSC phases of OKMS reprogramming. As for OKS reprogramming, this time the non-coding isoforms remained very lowly expressed throughout the process (**Supplementary Fig. 9g**). However, expression of the NMDex3 isoforms increased until the middle phase of reprogramming, regulating the proportion of functional isoforms. In later phases, the proportion of the NMDex3 isoforms decreased significantly (going from 38.1% in NPCs to 10.5% in the late phase), thus allowing an increase in proportion of the main protein-coding isoforms (going from 58.6% in NPCs to 88.8% in late phase) (**Fig. 5i**, **Supplementary Fig. 9f-g**, and **Supplementary Table 6**). Thus, in both reprogramming systems, isoforms leading to the proper SRSF2 protein dominate later in reprogramming through expression control of different isoforms, and more specifically for OKMS by IS of the non-coding isoforms to the benefit of the main coding isoforms. A similar effect could be observed when analyzing reprogramming from B cells with OSKM reprogramming factors (data from Stadhouders *et al.*[72]). Indeed, a significant increase in the proportion of isoforms leading to the proper SRSF2 protein can be seen throughout reprogramming, accompanied by a significant increase of the non-coding isoforms in the early phase, and their decrease in late reprogramming (**Supplementary Fig. 9h** and **Supplementary Table 6**). This suggests that this phenomenon is conserved across reprogramming systems from multiple differentiated cell types. To assess whether the variations in *Srsf2* isoform levels could have an impact on the role of this splicing regulator, we investigated the fate of known SRSF2 target gene transcripts[71] in our dataset (**Supplementary Table 6**). The first half of reprogramming was associated with an increase in non-functional isoforms (whether non-coding or NMD isoforms) for SRSF2 target mRNAs in both reprogramming systems (**Supplementary Fig. 9i**, **Supplementary Table 2**, and **Supplementary Data 4**), which was consistent with a diminished proportion of protein-coding isoforms. Conversely, increase of functional *Srsf2* isoforms in late reprogramming occurred in the same phase as the decrease in the non-functional isoforms of SRSF2 target mRNAs (**Fig. 5i** and **Supplementary Fig. 9i**). Overall, these results

suggest again that selective regulation through AS and isoform switching allows proper control of genes important for proper reprogramming and for the resulting iPSCs.

This regulation of genes by expression of non-coding isoforms is also consistent with changes in coding probability decreasing in the middle phase of reprogramming, to then increase in later phases as observed in our ΔPSI-based IS analysis (**Fig. 3d**). This also offers interesting parallels with the specific regulation of non-coding isoforms from both coding genes (**Fig. 5f-g**) and lncRNAs (**Fig. 5d**) observed during reprogramming and suggests that coding and non-coding isoforms are regulated differently and at different phases of reprogramming. With this in mind, and because we previously showed how during reprogramming IS can select for isoforms of protein-coding genes containing functional domains (**Fig. 4** and **Fig. 5i**), we reasoned that some lncRNAs undergoing IS may also harbor important functions that need to be regulated during the first half of the reprogramming process and thus for pluripotency acquisition.

Identification of a non-coding transcript regulating the pluripotent state

Because isoform switching of lncRNAs is under strict control during early reprogramming (**Fig. 5d**), we hypothesized that isoform switching could regulate several lncRNAs involved in the pluripotency regulatory network. To increase chances of identifying functionally relevant IS events in lncRNAs, we focused on lncRNAs undergoing isoform switching during NPCs to mid-reprogramming in either OKMS or OKS systems and that are identified in both our ΔPSI events-based analysis (n=1,657 lncRNA genes) and our DEIs-based analysis of (n=59 lncRNA genes) (**Fig. 6a**). We obtained 45 lncRNA candidates that were common to these two lists (**Fig. 6a** and **Supplementary Table 7**). In order to validate the importance of these candidates for reprogramming, we cross-referenced our list with lncRNAs that were reproducibly detected and differentially expressed in at least one other reprogramming system (i.e. OSKM or OKMS with 2° MEFs as a starting cell type) and in ESCs[16,25,41]. We identified ten candidates (**Fig. 6b-c**, **Supplementary Fig. 10a-b**, and **Supplementary Table 7**). From these, *2700038G22Rik*,

2010204K13Rik, *Gm11944*, and *Mir17hg* were the only four lncRNAs with expression increasing throughout both reprogramming datasets and remained consistently higher in iPSCs than in MEFs (**Fig. 6b-c** and **Supplementary Fig. 10a-b**). More specifically, the lncRNA *2700038G22Rik* showed this pattern consistently across multiple reprogramming datasets (**Fig. 6c**) suggesting a possible role in the acquisition of the pluripotent state. Moreover, a human syntenic ortholog could only be found for three of the four candidates: *2700038G22Rik, 2010204K13Rik* and *Mir17hg* lncRNAs (**Fig. 6e**, **Supplementary Fig. 10c**, and **Supplementary Table 7**). However, the human ortholog for *2010204K13Rik* lncRNA appeared to be the coding gene *PAGE4*, which was only expressed in late reprogramming (**Supplementary Fig. 10c**). For this reason, *2010204K13Rik* was not studied further. As for *Mir17hg*, it is a conserved precursor for the *miR-17-92* microRNAs cluster[73,74], which have been shown to be highly expressed in ESCs to regulate cell cycle progression[75] and expression of pluripotency genes[76,77]. They are also induced during the early phase of reprogramming[77,78] and enhance human fibroblast OKMS and OKS reprogramming[78,79] by decreasing TGFBR2 and p21 protein levels to overcome reprogramming barriers[79]. Identification of *Mir17hg* as a lncRNA whose by-products are well studied in ESCs and reprogramming confirmed that our identification pipeline is thus appropriate to discover lncRNAs with a function in pluripotency regulation, and suggested that the other top candidate *2700038G22Rik* could be important as well.

Interestingly, NPCs also express *2700038G22Rik* (**Fig. 6d**), and its expression increased even further during OKMS reprogramming but significantly decreased and then increased in the OKS system to reach iPSCs/ESCs levels (**Fig. 6d**). This increase in expression was reproducible in different datasets using 2° MEFs or 2° tail-tip fibroblasts, with expression levels in iPSCs comparable to those in control ESCs[16,18,41] (**Fig. 6c**). Moreover, expression of the human ortholog of *2700038G22Rik*, *SNHG26*, showed a similar expression pattern of up-regulation during human reprogramming[80] (**Fig. 6e**). *2700038G22Rik* lncRNA thus appeared as a promising regulator of pluripotency, as its expression was increasing during reprogramming, it is highly expressed in pluripotent stem cells, and has a conserved expression pattern in different mouse and human reprogramming systems.

Because *2700038G22Rik* (named *Snhg26* from here on) was identified as undergoing IS, we analyzed its different isoforms and identified 11 transcripts expressed among the 279,998 isoforms detected in our dataset (**Fig. 6f**). Only four could be identified in the latest Ensembl database version (release version 113.39, Aug. 2024)[81] (**Fig. 6f**), highlighting dataset specific isoform diversity and the importance of long-read data integration into public databases. In our reprogramming dataset, two isoforms showed differential expression between the early and middle phases. Those two isoforms, *G38722.216* ("*Snhg26.216*") and *G38722.217* ("*Snhg26.217*"), demonstrated an opposite expression pattern: *Snhg26.216* level was more elevated than the one of *Snhg26.217* in NPCs (**Supplementary Fig. 10d-e**), but decreased significantly in the middle phase of both OKMS and OKS reprogramming (**Fig. 6g** and **Supplementary Fig. 10d-e**). *Snhg26.217* expression increased drastically in OKMS reprogramming early phase (relative to NPCs) so that in middle phase, the proportion of *Snhg26.217* was significantly more (going from 13.2% in NPCs to 75.5% in middle phase) (**Fig. 6g**, **Supplementary Fig. 10d-e,** and **Supplementary Table 7**), thus leading to a clear isoform switching event. In OKS reprogramming, both *Snhg26.217* and *Snhg26.216* expression decreased, but the fraction of *Snhg26.217* isoform increased in early reprogramming (going from 13.2% in NPCs to 25.7% in early phase) at the expense of the *Snhg26.216* isoform (**Fig. 6g** and **Supplementary Table 7**). The strong increase of *Snhg26.217* over *Snhg26.216* isoform was further validated by RT-PCR followed by lrRNAseq (**Supplementary Fig. 11a** and **Supplementary Data 5**). Similarly, in OSKM reprogramming from B cells[72], opposed expression dynamics were observed for *Snhg26.217* and *Snhg26.216* (**Supplementary Fig. 11b**). Early reprogramming from B cells is associated with an increase in the proportion of *Snhg26.216* isoforms, at the expense of *Snhg26.217* expression. Later in reprogramming, *Snhg26.217* levels rise back to their original levels, suggesting that *Snhg26* isoform switching is conserved across multiple reprogramming systems.

To assess whether *Snhg26* isoforms are associated with reprogramming outcome, we analyzed single-cell RNA sequencing data from reprogramming cells[18] (**Supplementary Fig. 11c** and **Supplementary Table 6**). During reprogramming (i.e. "transitional reprogramming cells"), 23% of cells

express only *Snhg26.217*, 11% express only *Snhg26.216*, and 7% express both, and the proportion of *Snhg26.217* expressing cells increased significantly when cells reach the pluripotent state (**Fig. 6h**). Moreover, transitional cells expressing *Snhg26.217* isoform have significantly higher expression of ESC-like genes when compared to cells expressing *Snhg26.216* (**Fig. 6i** and **Supplementary Fig. 11d**). This mirrors the subset of transitional reprogramming cells expressing known markers associated with favorable reprogramming outcome (e.g. *Dppa2*, *Utf1* and *Esrrb*[48]) (**Fig. 6j** and **Supplementary Fig. 11e-f**). This data, as well as *Snhg26* conserved expression pattern among different reprogramming systems and species, suggests that this lncRNA, more specifically the *Snhg26.217* isoform, may function as an important regulator of the early stages of pluripotency induction.

Snhg26.217 **lncRNA enhances the reprogramming route to pluripotency**

In order to investigate if *Snhg26*, more specifically isoform *217*, could have a function during the acquisition of mouse pluripotency, we performed reprogramming experiments where we overexpressed either a control RNA or the *Snhg26.217* lncRNA isoform with the reprogramming factors OKS (**Supplementary Table 7**). Upon DOX addition, the proportion of reprogramming mCherry+ cells peaked earlier for OKS + *Snhg26.217* (80% of cells at D8) than for OKS + control (58.35% of cells at D14) (**Fig. 7a**, **Supplementary Fig. 1**, and **Supplementary Fig. 12a-b**). However, activation of the *Oct4* locus (i.e. GFP+ cells) was similar starting from D14 (2.12% for OKS + control; 1.44% for OKS + *Snhg26.217*) and increasing similarly between the conditions to reach their maximum at D20 (21.78% for OKS + control; 18.50% for OKS + *Snhg26.217*) (**Fig. 7b**). By this time, most cells had spontaneously silenced transgenes expression (11.05% of mCherry+ cells for OKS + control; 13.39% for OKS + *Snhg26.217* at D20) (**Fig. 7a**). Surprisingly, the increased number of mCherry+ cells when *Snhg26.217* was expressed did not lead to significantly more GFP+ cells upon DOX removal (63.35% for OKS + control; 75.55% for OKS + *Snhg26.217* at D14) (**Fig. 7b**), even if mCherry levels remained high (going from 51.61% at D6 to 73.4% at D10 for OKS + *Snhg26.217,* compared to 37.70% at D10 for OKS +

control) (**Fig. 7a**). For both conditions, most cells silenced the exogenous reprogramming factors upon DOX removal, starting by D14 (5.72% for OKS + control; 7.07% for OKS + *Snhg26.217*) (**Fig. 7a**). The fact that the proportion of mCherry+ cells for OKS + *Snhg26.217* was more than twice the proportion of OKS + control reprogramming cells, and that this was not reflected in the percentage of GFP+ cells emerging, suggested two scenarios which could explain how *Snhg26.217* expression increased the number of reprogramming cells: 1) *Snhg26.217* could directly affect proliferation of cells that commit to reprogramming but not the establishment of iPSCs (e.g. cells are not able to complete reprogramming to the final iPSC stage, possibly remaining blocked at the pre-iPSC stage), or 2) *Snhg26.217* may influence the increase in iPSC colony formation (i.e. more cells commit to iPSCs) (**Fig. 7c**). For the latter scenario, the total number of cells would be greater under *Snhg26.217* expression because of increase in reprogramming cells (i.e. mCherry+ cells); thus, looking at percentages will not reflect this increase as global cell confluency is not considered in such calculations. For example, an increase in mCherry+ cells leads to an increase in the number of cells that commit to reprogramming; however, for cells to become iPSCs (i.e. GFP+ cells) and for reprogramming to complete, transgene expression must be silenced. This will lead to a mixed population of GFP+ and negative cells, which will show an increase in cell density but not necessarily the proportion of iPSCs (please refer to **Supplementary Note 2** for details).

To test these two scenarios, we first addressed whether *Snhg26.217* expression could increase the proliferation of reprogramming cells. We found that *Snhg26.217* expression significantly and consistently increased the population of mCherry+ cells (297,253 cells for OKS + control; 471,693 cells for OKS + *Snhg26.217* at 120 hours post-DOX induction) (**Fig. 7d**) and the total number of cells (**Supplementary Fig. 12c**). However, no significant difference was observed in mean doubling time of these mCherry+ or total populations between 24h and 120h of DOX induction between control and *Snhg26.217* conditions (**Fig. 7d** and **Supplementary Fig. 12c**), suggesting that *Snhg26.217* expression does not influence cell proliferation. We also quantified the doubling times of the starting NPC lines following the expression of a control or *Snhg26.217* RNA, in NPC medium and independent of reprogramming medium conditions. Once again, their average proliferation rates were similar (**Supplementary Fig. 12d**), suggesting that

Snhg26.217 overexpression did not impact the growth of the starting NPC lines. Finally, to make sure that *Snhg26.217* expression does not influence the division rate of pluripotent cells themselves, we compared the doubling times of ESCs overexpressing *Snhg26.217* or a control RNA in 2i-LIF + DOX medium every 12 hours. Again, there was no difference in the proliferation curves of ESCs when comparing control to *Snhg26.217* overexpression (**Supplementary Fig. 12e**). These results demonstrated that *Snhg26.217* did not affect cell proliferation and that scenario 2 (i.e. increase in iPSC colony formation) was more plausible.

To address the second scenario, and since the increased number of reprogramming cells led to a global cell population increase, it was thus important to account for the global density of cells in the calculation of mCherry+ and GFP+ cells during reprogramming to properly measure the *Snhg26.217* effect on the process (please refer to **Supplementary Note 2** for details). By doing this, we observed that the mean density of mCherry+ cells progressively increased to reach its maximum at D14 for both OKS + control (2,065,072 cells/mL) and OKS + *Snhg26.217* (3,823,884 cells/mL) conditions and subsequently decreased to give rise to GFP+ cells (**Fig. 7e-f**). Upon DOX removal, transgene expression dropped rapidly between D10 and D14 for both conditions, and we observed a strong induction of iPSCs at D14, with 16 times more GFP+ cells than in control reprogramming (112,328 cells/mL for OKS + control; 1,877,379 cells/mL for OKS + *Snhg26.217*) (**Fig. 7f**). These results highlight that scenario 2 is the most plausible explanation for our observations and that *Snhg26.217* may accelerate reprogramming to iPSCs. Therefore, by analyzing cell density, it allowed us to visualize that the number of both reprogramming cells and iPSCs were enhanced by *Snhg26.217* overexpression compared to the control condition. To validate the robustness of this result, we analyzed the density of GFP+ cells using other reprogramming strategies. We found that the density of GFP+ cells was always higher when *Snhg26.217* was expressed compared to the control condition regardless of reprogramming factors order, cell origin, or whether it was done using transient or stable transfections (**Fig. 7g** and **Supplementary Fig. 13a-f**). Together, these results validated that *Snhg26.217* significantly enhanced the induction of iPSCs in various reprogramming systems.

We next asked if *Snhg26.217* could influence the frequency of cells able to engage into reprogramming and to reach the iPSC stage by changing cell identity and enhancing the cells' rate of reprogramming. To assess this, we performed a limiting dilution assay, a robust assay usually used to assess the frequency of stem cells in a cell population *in vitro* (please refer to **Supplementary Note 3** for details)[82-84]. We used the same OKS reprogramming system as before, but this time we seeded the starting cell lines as graded dilutions (5, 25, 125 or 500 cells/well) with 24 replicates in a culture plate. We induced the transgenes with doxycycline and calculated the percentage of wells negative of any reprogramming events at D12 and then at D27 after removing DOX at D18. We found that in the control condition, about 0.6% of cells were able to give rise to iPSCs during OKS reprogramming (0.60% (1/178 cells) at D12; 0.57% (1/181 cell) at D27) (**Fig. 7h-i**). However, when we overexpressed *Snhg26.217* along with the reprogramming factors, this frequency increased to about 1% (0.89% (1/114 cells) at D12; 1.08% (1/95 cells) at D27) (**Fig. 7h-i**). Together, these results led us to conclude that *Snhg26.217* accelerates the reprogramming kinetics by enhancing the overall efficiency and increasing cell state changes rather than increasing cell division (for a more detailed explanation, see **Supplementary Note 2**).

To validate that iPSCs generated with *Snhg26.217* are *bona fide* iPSCs, we isolated clones from the OKS + *Snhg26.217* reprogramming experiments and tested their contribution to the formation of chimeras using diploid aggregations. E8-8.5 embryos showed high contribution for OKS + *Snhg26.217* iPSCs, as visualized with the negative cells/area in tdTomato-positive embryos (**Supplementary Fig. 13g**). As expected, iPSC contribution was also limited to the embryonic area (i.e. epiblast derivatives), with extraembryonic tissues remaining tdTomato-positive. When aggregated to albino mice, the iPSC clones contributed 5-55% to chimera formation, as judged by coat color (**Fig. 7j** and **Supplementary Fig. 13h**). Finally, OKS + *Snhg26.217* iPSCs also successfully formed teratoma when injected into mice, and cells derived from the three germ layers were represented to the same extent as OKS + control iPSCs (**Supplementary Fig. 14**). This confirmed that *Snhg26.217* expression during reprogramming gave rise to *bona fide* iPSCs.

1 ***Snhg26* lncRNA is a developmental gene important for the pluripotent state**

2 To further investigate the regulation and function of *Snhg26* lncRNA in pluripotency, we used our
3 previous ChIPseq during reprogramming of 2˚MEFs to iPSCs[16] and analyzed the chromatin landscape
4 around *Snhg26* locus. The *Snhg26* promoter was marked by H3K4me3 and H3K27me3 and devoid of
5 DNA methylation in 2˚MEFs, but lost the H3K27me3 mark as cells progressed through reprogramming
6 towards iPSCs (**Fig. 8a**), suggesting a bivalent locus that is activated upon loss of the repressive mark
7 H3K27me3. This is in accordance with up-regulation of *Snhg26* during reprogramming and a sign of a
8 developmentally regulated genes[85,86]. Moreover, a previous study identified *Snhg26* as part of a cohort of
9 ESC-specific lncRNAs and showed that knock-downs of KLF2 or N-MYC pluripotency-associated
10 transcription factors, both lead to a down-regulation of *Snhg26* (named *linc1533* in that study) in mouse
11 ESCs[25]. They also showed the binding of KLF4, C-MYC and N-MYC to *Snhg26* promoter[25], suggesting
12 that *Snhg26* may be regulated by the pluripotency-transcriptional network.

13 Because those results suggested a role for *Snhg26* in ESCs, we wondered if *Snhg26* expression could
14 be conserved in the broader context of pluripotency and development. To this end, we evaluated total
15 *Snhg26* expression during mouse early embryo development using a published dataset[13]. Interestingly,
16 *Snhg26* expression peaks at the 4-cell stage of pre-implantation embryos, and remains expressed up to the
17 blastocyst stage (i.e. pluripotent state), albeit at a lower expression level (**Fig. 8b**). This peak expression
18 corresponds to the zygotic genome activation stages in mice and suggests a role for *Snhg26* in embryo
19 development. Furthermore, if we examine *Snhg26* expression during *in vitro* differentiation of ESCs
20 towards cortical neurons[87], we observe that it drops drastically from the first days of the differentiation
21 process (**Supplementary Fig. 15a**). These results suggest that expression of *Snhg26* is specific to early
22 embryo development and the associated pluripotent state.

23 To address its role in pluripotency, we first sought to determine where *Snhg26* lncRNA is localized as
24 localization of a lncRNA is generally indicative of its function[88,89]. We found *Snhg26* to be nuclear
25 (83.8% on average) according to quantifications performed after cellular fractionation of mouse ESCs

(**Supplementary Fig. 15b** and **Supplementary Table 8**). Its expression was similar to other well-characterized nuclear lncRNAs such as *Xist* (88.2% on average), *Malat1* (85.4% on average) and *Neat1* (89.5% on average). Moreover, *in situ* hybridization of probes targeting *Snhg26* transcripts demonstrated strong co-localization with DAPI staining (**Supplementary Fig. 15c**), further validating *Snhg26* nuclear localization in mouse ESCs. Therefore, we knocked-down *Snhg26*'s transcription in mouse ESCs by targeting its promoter with a CRISPR-interference (CRISPRi) system[90,91], which also ensures that all its transcripts were targeted. Knock-down of *Snhg26*'s transcription for 5 days in mouse ESCs led to down-regulation of pluripotency genes *Oct4, Nanog* and *Tfcp2l1* (**Fig. 8c**), confirming that *Snhg26* was involved in the pluripotency regulatory network of ESCs.

Finally, in order to link the role of the specific *Snhg26.217* isoform in enhancing reprogramming with the function of the *Snhg26* gene in maintenance of pluripotency, we asked whether *Snhg26* isoforms had similar or divergent roles during early embryonic development. First, we assessed their expression in a mouse embryo published dataset[13] and found that while *Snhg26.216* was expressed between the 2-cell and blastocyst stages, *Snhg26.217* expression levels were always higher, particularly high at the 2-cell and 4-cell stages (**Fig. 8d** and **Supplementary Table 8**). This suggested that *Snhg26.217* may have a more important role in early development than *Snhg26.216*. To assess this, we generated a *Snhg26* knock-out (KO) mouse ESC line by removing the last exon of *Snhg26*, which represents 92.5% of its whole exonic sequence using CRISPR-Cas9. We were able to generate an $Snhg26^{KO}$ ESC clone, that was then engineered to overexpress either *Snhg26.217* or *Snhg26.216* transcripts using DOX induction. Three-dimension blastoid models, *in vitro* models of pre-implantation blastocysts, were derived from the KO + *Snhg26* transcript lines, along with wild-type cells, over three days of DOX induction and stained for markers of the inner cell mass (SOX2) and the primitive endoderm (GATA6). We found that *Snhg26* KO led to smaller blastoids (**Fig. 8e-f** and **Supplementary Fig. 15d**), with a significantly reduced number of cells (**Fig. 8g**) compared to wild-type ones. Moreover, they displayed less SOX2-positive cells and an unchanged number of GATA6-positive cells, suggesting that lack of *Snhg26* lncRNA prevented the formation of inner cell mass-like cells (**Fig. 8h**). Strikingly, overexpression of *Snhg26.217* isoform

rescued the *Snhg26*^{KO} blastoid size and total number of cells, even increasing those features (**Fig. 8f-g** and **Supplementary Fig. 15d**). This time, the number of SOX2 and GATA6-positive cells was not significantly affected (**Fig. 8e-h**), suggesting that *Snhg26.217* isoform rescues development of inner cell mass-like cells. On the other hand, *Snhg26*^{KO} blastoids overexpressing *Snhg26.216* isoform displayed a phenotype similar to the *Snhg26*^{KO} blastoids (**Fig. 8e-h** and **Supplementary Fig. 15d**). These results suggest that *Snhg26.217* isoform expression, and not that of *Snhg26.216,* is important for proper embryo development and formation of the inner cell mass of mouse early blastocysts.

Altogether, the effects of overexpression, knock-down, and knock-out of *Snhg26* and its isoforms in pluripotency models such as reprogramming, ESCs and blastoids demonstrate the role of *Snhg26* lncRNA, and more specifically of *Snhg26.217* isoform, in the instauration and maintenance of the pluripotent state. Thus, lrRNAseq analysis of our reprogramming systems enabled us to identify a functional isoform of a new lncRNA involved in regulating the pluripotent state.

SNHG26 lncRNA has a conserved function in human pluripotency

Because *Snhg26* was selected from a lncRNA candidate list for its conserved expression pattern during mouse and human reprogramming, we wondered if it could have a similar function in the regulation of human pluripotency. To this end, we assessed the expression of its human counterpart, *SNHG26* lncRNA, during human embryo development using a published dataset[92]. Similarly to the mouse lncRNA, *SNHG26* expression started at the 8-cell stage of pre-implantation embryos, which corresponds to the zygotic genome activation stages in humans (**Fig. 9a**). *SNHG26* remained quite highly expressed in human pluripotent stem cells. Because mouse and human ESCs (hESCs) correspond to different states of pluripotency[93,94], we examined *SNHG26* expression in primed and *naïve* hESCs, as well as *in vitro* equivalent of 8 cell-like cells[95]. We found that *SNHG26* was highly expressed in *naïve* and 8 cell-like cells compared to primed hESCs (**Fig. 9b**), which is in keeping with results from pre-implantation embryo datasets. Finally, exploring *SNHG26* expression during *in vitro* differentiation of hESCs towards cortical neurons[96], we observed that its expression dropped drastically from the beginning of the process

(**Supplementary Fig. 16a**). These expression patterns suggest that *Snhg26* may have a specific and conserved role in early embryo development and the associated pluripotent state.

As with mouse ESCs, we knocked-down *SNHG26*'s transcription in hESCs by CRISPRi. Contrary to results in mouse ESCs, in hESCs, no significant decrease was observed for the core pluripotency genes *OCT4* and *NANOG* (**Supplementary Fig. 16b**). This can be explained by the fact that, as previously mentioned, mouse and human ESCs do not correspond to the same state of pluripotency, and *SNHG26* is more expressed in the *naïve* state. However, the two markers of primed pluripotency *SALL4* and *OTX2* were significantly down-regulated (24% and 33%, respectively). Furthermore, expression of some *naïve* pluripotency markers was even more decreased (40%, 32% and 42% for *TFCP2L1*, *REX1* and *DPPA4*, respectively) (**Supplementary Fig. 16b**). Deeper look at how overall gene expression was affected by *SNHG26* knock-down in hESCs by lrRNAseq revealed that additional markers of pluripotency were down-regulated, such as *DPPA3, UTF1* and *FGF5* (**Fig. 9c-d** and **Supplementary Table 8**). These results suggested that similar to its mouse ortholog, *SNHG26* was important to maintain the pluripotent state of hESCs. Interestingly, GO-term enrichment analysis showed that genes up-regulated following *SNHG26* knock-down were linked to regulation of gene transcription and were tumor suppressor genes (**Fig. 9e** and **Supplementary Table 8**). This suggested that *SNHG26* lncRNA could act as a transcriptional regulator in hESCs, in keeping with its nuclear localization in mouse ESCs. On the opposite, down-regulated genes were associated to cell migration, cancer-related pathways and cancer resistance (**Fig. 9f** and **Supplementary Table 8**). This is interesting as *SNHG26* lncRNA has been shown to be up-regulated in many types of resistant cancers[97-101], and genes from the pluripotency gene network are overexpressed in cancer, promoting stem cell-like properties and uncontrolled growth, thus driving tumor initiation, progression and resistance to treatment[102]. This suggests that *SNHG26* could have a role in tumorigenesis by promoting stem cell-like properties. Consistent with this notion, *SNHG26* has been shown to be a pivotal regulator of keratinocyte progenitors and their transition from an inflammatory phase to a subsequent proliferative phase during both mouse and human wound healing[103].

Next, in order to validate the conserved role of *SNHG26* lncRNA in regulation of pluripotency, we used a pool of antisense oligonucleotides (ASOs) to knock-down *SNHG26* during reprogramming of human dermal fibroblasts to iPSCs (**Fig. 9g**). ASOs were employed due to the technical challenges associated with generating CRISPRi lines from fibroblasts. Transfection of ASOs throughout reprogramming resulted in a 70% reduction in the expression of *SNHG26* by day 12, relative to cells treated with non-targeting control ASOs (**Supplementary Fig. 16c**). Although no morphological differences were observed at the end of reprogramming factors induction (i.e. D10), iPSC colonies with *SNHG26* knock-down showed a lower density and size compared to reprogramming without and with control ASOs (**Supplementary Fig. 16d**). Quantification of NANOG, OCT4, and SOX2-positive colonies at D20 of reprogramming also demonstrated a reduction in the number of colonies following *SNHG26* knock-down, with noticeably smaller sizes than the control conditions (**Fig. 9h-i** and **Supplementary Fig. 16e**). These results demonstrated that *SNHG26* lncRNA has a conserved role in the acquisition of the pluripotent state, and that the function of this lncRNA is not limited to reprogramming of neural progenitor cells but also to fibroblasts.

Next, we asked whether we could find specific isoforms that mediate the function of human *SNHG26*. *SNHG26* has 22 human isoforms according to Ensembl database (release version 107, July 2022)[104], with only 10 being expressed in human PSCs or early embryo datasets (**Supplementary Fig. 17** and **Supplementary Table 8**)[92,95,96,105-109]. We identified 4 that have a conserved exon-intron structure with the mouse *Snhg26.217*, and 4 resembling the *Snhg26.216* isoform (**Supplementary Fig. 17**). Among the isoforms resembling *Snhg26.217, SNHG26.201* has already been shown to regulate cell fate and proliferation of keratinocytes[103]. However, when looking at a dataset from early human embryonic development samples, only 6 of the resembling transcripts were detected, and *SNHG26.201* was lowly expressed (**Fig. 9j**). Strikingly, only *SNHG26.219,* which has a conserved structure with *Snhg26.217,* had the same expression pattern as its mouse ortholog, being highly expressed at the 8-cell stage (**Fig. 9j**). Moreover, *SNHG26.219* was the transcript the most expressed in *in vitro* 8 cell-like cells (**Fig. 9k**). These

results suggest that *SNHG26.219* transcript may be the human ortholog of *Snhg26.217* and possess a similar function in the regulation of pluripotency. Further work is required to establish if this is the case.

Thus, the genomic conservation and identical expression profiles over several biological systems suggested that *Snhg26* lncRNA could share a conserved functional role in pluripotency regulation (refer to **Fig. 10** for a working model). Thus, lrRNAseq analysis of our optimized reprogramming systems enabled us to identify a functional isoform of a new lncRNA whose function is conserved in regulating the pluripotent state.

Discussion

Previous studies have demonstrated that AS regulation is an important factor in cell fate determination, for instance during somatic cell reprogramming to pluripotency. Two reports[32,35], using different reprogramming systems, have explored genome-wide AS across different timepoints of reprogramming and demonstrated that specific splicing patterns and regulators are linked to reprogramming. Taking advantage of full-length transcripts generated by long-read RNAseq, we explored this further and undertook the first lrRNAseq-mediated characterization of isoform diversity during the reprogramming of cells towards an induced pluripotent state. Our work identifies for the first time that isoform diversity and switching are highly implicated in the process. We also show that this is not limited to protein-coding RNA isoforms, but rather non-coding RNA isoform switching plays a major role in regulating gene function and reprogramming outcome. In fact, among genes that are affected by this IS control mechanism is *Srsf2*, which was also identified by these previous studies and others[29,31] as an important AS regulator during reprogramming and pluripotency. Expression of non-coding isoforms was down-regulated and made way for expression of isoforms that lead to functional proteins. We also show that the splicing and isoform expression patterns of these regulators differs between our reprogramming systems and may affect their functional outcome. Both OKMS and OKS reprogramming systems have the same start (NPCs) and end (iPSCs) points but use different interplay between isoform expression and AS to achieve the same outcome. This also suggests that isoform diversification and switching promote

functional outcomes that are favorable for reprogramming and exposes a level of isoform regulation that has not been shown before. Thus, in our study, we aimed at identifying such outcomes and their effect on the regulation of the pluripotent state.

We began by first devising an appropriate reprogramming system model that recapitulates known features of reprogramming in terms of timelines, gene expression profile and alternative splicing events. Running lrRNAseq at different timepoints of reprogramming, we brought to light the underestimated diversity of isoforms expressed during this process, with about 65% of the total detected isoforms being previously unannotated and many originating from new splice junctions (i.e. "Novel-Not-in-Catalog" transcripts). This reinforces the utility of lrRNAseq to address the shortcomings of short-read RNAseq in identifying full-length isoforms. Closer examination of isoforms showed that the transcripts that were classified as non-coding were regulated in the early and middle phases of both reprogramming systems. We found that most differentially abundant isoforms were at the transition to the early phase of reprogramming. By the middle phase of the process, a switch was observed in the abundance of alternatively spliced isoforms. We also found that less abundant spliced isoforms had a lower coding potential before this switch, and this hinted at a pressure to regulate specifically non-coding isoforms in the early phases of reprogramming. This demonstrates that non-coding isoforms undergoing isoform switch may have important functions during acquisition of pluripotency. These findings show that isoform diversification and switching of non-coding genes, such as lncRNAs, add a new level of regulation in pluripotency establishment.

Previous research has shown involvement of lncRNAs in pluripotency regulation[25,110-112], but only a few examples of functionally tested lncRNAs exist. Moreover, unlike protein-coding genes, which give rise to a functional product, it is often difficult to identify the functional isoform and the mechanism of action of a lncRNA. Our study provides a new way of identifying functionally relevant lncRNAs in the context of pluripotency. By integrating our data with several datasets on reprogramming and pluripotency, we identified a lncRNA that undergoes isoform switch during early reprogramming. The overexpression of the *Snhg26.217* isoform greatly increased the number of reprogramming cells and *bona*

fide iPSCs. Furthermore, knock-down of human *SNHG26* resulted in a less efficient reprogramming. *Snhg26* in mouse and human ESCs down-regulated expression of several pluripotency markers, validating the conserved role for this new lncRNA in the regulation of pluripotency. Moreover, the conserved expression pattern of mouse and human *Snhg26* during embryo development and their common role in maintaining the pluripotent state is of interest, as it illustrates biological contexts in which the *Snhg26* lncRNA is evolutionary conserved. Indeed, *SNHG26* knock-down in hESCs up-regulated several tumor suppressor genes, linking its role in promoting stem cell features with several studies that showed its association with multiple cancer types. Finally, a recent study[103] demonstrated the conserved function of *SNHG*26 in impeding human and mouse keratinocytes inflammatory response and promoting their progenitor proliferation and migration in order to re-epithelialize skin wounds. It would thus be of interest to investigate which isoform of the human *SNHG26* is involved in hESC maintenance, and if it is the same as during keratinocytes wound recolonization. This could further bring to light the importance of alternative splicing in cell fate determination.

One of the limitations of this study is the use of a reprogramming system where transgenes are randomly integrated within the genome of the reprogramming cells. We can thus not exclude that this parameter could further affect the highlighted differences between OKMS and OKS reprogramming, nor the resulting enhanced reprogramming to iPSCs when *Snhg26* lncRNA is expressed. However, our system is based on a heterogenous cell population, and the effects of *Snhg26* on reprogramming kinetics were validated multiple times using different cell lines and reprogramming strategies.

In summary, it is important to note that only by analyzing isoform switching of *Snhg26* we were able to identify a functional isoform of this lncRNA. Given the current state of lncRNA annotations, which is highly complex and of unknown functional consequence, our study highlights the benefits of using lrRNAseq to recognize isoform diversity and switching and how in-depth analysis of such processes is key to identifying functionally relevant lncRNA isoforms. The dataset we generated also stands as a valuable resource for other researchers to study the role of alternative splicing and isoform diversity in the regulation of cell fate.

Methods

Generation of the destination vectors

For reprogramming, the *piggyBac* (PB) transposon vectors PB-TAC-OKMS (Addgene #80480), PB-TAC-OKS (OKS sequence was amplified from Addgene #19771 and reverse Gateway™ cloned into PB-TAC-OKMS to replace the OKMS sequence), and PB-TAC-OSKM (Addgene #80481) containing a doxycycline-inducible TetO promoter, as well as a polycistronic reprogramming cassette with either *Oct4*, *Klf4*, *c-Myc* and *Sox2* genes (OKMS), *Oct4*, *Klf4* and *Sox2* genes (OKS), or *Oct4, Sox2, Klf4* and *c-Myc* (OSKM) and an *mCherry* reporter gene were kindly gifted by Dr. Andras Nagy (University of Toronto). The pCMV-hyPBase plasmid[113] from the Sanger Institute was used to express the PB transposase.
For transgene overexpression, the PB transposon destination vectors PB-TetO-AIO carrying our genes of interest (e.g. *Snhg26*) were generated by digesting an in-house PB-TetO-Gateway plasmid and the PB-TAC-ERP2 (Addgene #80478) or PB-TAC-ERN (Addgene #80475) plasmids with MluI-HF® (New England Biolabs (NEB) #R3198) and SbfI-HF® (NEB #R3642), followed by ligation using a T4 DNA ligase (ThermoFisher Scientific #EL0016) according to the manufacturer's instructions. They carry a doxycycline-inducible TetO promoter, a Gateway™ cloning cassette as well as the constitutive reverse tetracycline-controlled TransActivator (rtTA) and a puromycin (P) or neomycin (N) resistance gene.
For CRISPRi-mediated knock-downs, an in-house dCas9-BFP-KRAB construct was PCR-amplified from the pHR-SFFV-dCas9-BFP-KRAB plasmid (Addgene #46911) using the Deoxynucleotide (dNTP) Solution Set (NEB #N0446S) and the Q5® High-Fidelity polymerase (NEB #M0491) according to the manufacturer's instructions. The fragment was then inserted into the PB-TetO-AIO destination vector using Gateway™ cloning (ThermoFisher Scientific #11791020) strategy according to the manufacturer's instructions. The sequences of all plasmids were verified by Sanger sequencing and inducible expression validated by RT-qPCR. Complete sequences of all plasmids are available upon request.

Cloning of the *Snhg26* and control transcripts into destination vectors

Total RNA was extracted from mouse embryonic stem cells (ESCs) and reversed-transcribed into cDNA using SuperScript™ III Reverse Transcriptase (ThermoFisher Scientific #18080044) according to the manufacturer's instructions. Primers designed to target *Snhg26* transcript 217 (*Ensmust00000181574* or "*Snhg26.217*") were used with the dNTP Solution Set (NEB #N0446S) and the Q5® High-Fidelity polymerase (NEB #M0491) according to the manufacturer's instructions. *Snhg26* PCR-amplified cDNA transcript sequence was validated by Sanger sequencing and cloned into the PB-TetO-AIO destination vector using Gateway™ cloning (ThermoFisher Scientific #11789020 & #11791020) strategy according to the manufacturer's instructions. For *Snhg26* transcript 216 (*Ensmust00000268635* or "*Snhg26.216*"), the *Snhg26.217* containing plasmid was digested with BglII (NEB #R0144) and XcmI (NEB #R0533) to remove a part of exon 4. Primers to ligate back the plasmid were designed so that they also contain a missing part of the truncated exon 4 sequence. Those were added by ligation using a T4 DNA ligase (ThermoFisher Scientific #EL0016) according to the manufacturer's instructions. Similarly to *Snhg26.217*, a cDNA antisense to the *Luciferase* gene (control condition) was PCR-amplified using primers designed to target the pGL3-Basic plasmid (Promega #E1751) and cloned by Gateway™ cloning as described above. The sequences of all plasmids were verified by Sanger sequencing and inducible expression validated by RT-qPCR. Complete sequences of the resulting plasmids are available upon request (primer sequences are listed in **Supplementary Table 7**).

Mouse ESC culture

ESC lines, including R1 (RRID: CVCL_2167) and Oct4-GFP (harboring a *GFP* reporter gene under the control of the *Oct4* gene promoter from Viswanathan *et al*.[114]) (both gifted by Dr. Andras Nagy, University of Toronto) were routinely maintained in a humidified incubator at 37°C under 5% CO2. Cells were propagated in 2i-LIF medium on 0.1% porcine gelatin (Sigma-Aldrich #G1890). 2i-LIF medium consists of 20ng/mL LIF (home-made from pGEX-2T-MLIF[115]), 3μM CHIR-99021 (Selleck Chemicals #S2924) and 1μM PD0325901 (Selleck Chemicals #S1036) in N2B27 media. N2B27 media consists of a 1:1 mixture of DMEM/F-12 media mix (ThermoFisher Scientific #11330057) supplemented with 1X N-2

supplement (ThermoFisher Scientific #17502048) and 20µg/mL human recombinant Insulin (ThermoFisher Scientific #12585014), and Neurobasal™ media (ThermoFisher Scientific #21103049) supplemented with 1X B-27™ supplement (ThermoFisher Scientific #17504044), 1mM sodium pyruvate (ThermoFisher Scientific #11360070), 2mM L-glutamine (ThermoFisher Scientific #35050061), 1X MEM Non-Essential Amino Acid solution (ThermoFisher Scientific #11140050) and 10^{-4}M 2-Mercaptoethanol (ThermoFisher Scientific #21985023). Medium was refreshed every other day and ESCs were passaged every 2-3 days. For passaging, cells were treated with TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) and passaged as single-cells (usually 1/10 ratio). All cell lines were negative for mycoplasma.

Generation of Oct4-GFP neural progenitor cell (NPC) line

Oct4-GFP ESCs were dissociated into single-cells with TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) and seeded on 0.1% porcine gelatin (Sigma-Aldrich #G1890) in FBS-LIF medium. FBS-LIF (FBSL) medium consists of DMEM (high glucose, ThermoFisher Scientific #11965118) supplemented with 15% Fetal Bovine Serum (FBS, ThermoFisher Scientific #12483020), 1mM sodium pyruvate (ThermoFisher Scientific #11360070), 2mM L-glutamine (ThermoFisher Scientific #35050061), 1X MEM Non-Essential Amino Acid solution (ThermoFisher Scientific #11140050), 10^{-4}M 2-Mercaptoethanol (ThermoFisher Scientific #21985023) and LIF at 20ng/mL. The resulting primed cells were passaged at day 2 (D2) and day 4 in FBSL medium on 0.1% porcine gelatin using TrypLE™ Express Enzyme. On D4, cells were seeded at a density of 1 x 10^3 cells/cm^2 in FBSL medium on 0.1% porcine gelatin. Twenty-four hours later, medium was switched to N2B27 medium with 35µg/mL human recombinant Insulin (ThermoFisher Scientific #12585014) and replaced every other day. At D10, cells were dissociated with TrypLE™ Express Enzyme and transferred as single-cells into Corning™ Costar™ Ultra-Low Attachment microwell plates (Fisher Scientific #07-200-601) in Selection medium (2:1 mixture of N2B27 medium with 35µg/mL human recombinant Insulin supplemented with 20ng/mL human recombinant EGF (Peprotech #AF-100-15), 20ng/mL human recombinant FGF-2 (Peprotech

#100-18B) and 500ng/mL heparin sodium salt (Sigma-Aldrich #H3393)) to promote neurosphere generation. Medium was replaced every other day and the neurospheres were passaged every 3-4 days with TrypLE™ Express Enzyme until D20. Dissociated neurospheres were then seeded in NPC medium on Geltrex™-coated plates (hESC-qualified, growth factor-reduced, ThermoFisher Scientific, #A1413302). NPC medium consists of N2B27 supplemented with 20ng/mL human recombinant EGF, 20ng/mL human recombinant FGF-2 and 500ng/mL heparin sodium salt. Medium was refreshed every other day and NPCs were passaged every 2-3 days. For passaging, cells were treated with StemPro™ Accutase™ Cell Dissociation Reagent (ThermoFisher Scientific #A1110501) and passaged as single-cells (usually 1/6-1/8 ratio). All cell lines were maintained in a humidified incubator at 37°C under 5% CO2 and were negative for mycoplasma.

Human embryonic stem cells (hESCs) culture

Primed hESCs line H9 (RRID: CVCL_9773) were routinely maintained in a humidified incubator at 37°C under 5% CO2. Cells were propagated in mTeSR™1 medium (StemCell Technologies #85850) on Geltrex™-coated plates (hESC-qualified, growth factor-reduced, ThermoFisher Scientific #A1413302). Medium was refreshed every day and hESCs were passaged every 4-5 days. For passaging, cells were treated with Dispase (StemCell Technologies #07923) and passaged as small clumps using a Corning™ Cell Lifter (Fisher Scientific #07-200-364) (usually 1/10 ratio). All cell lines were negative for mycoplasma. Human ESCs were obtained commercially, and all experiments pertaining to hESCs were approved by the Stem Cell Oversight Committee as outlined by Canadian Institutes of Health Research as per the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2) and the Agreement on the Administration of Agency Grants and Awards by Research Institutions.

Transfection of mouse ESCs and NPCs

ESC lines were seeded at a density of 13 x 10^3 cells/cm^2 on plates coated with 0.1% porcine gelatin (Sigma-Aldrich #G1890) in 2i-LIF medium. Alternatively, NPC lines were seeded at a density of 30 x 10^3

cells/cm^2 on Geltrex™-coated plates (hESC-qualified, growth factor-reduced, ThermoFisher Scientific, #A1413302) in NPC medium.

The following day, cells were transfected with 200-500ng of hyPBase transposase and 1.8-2µg of PB-TetO-AIO constructs using Lipofectamine™ Stem Transfection Reagent (ThermoFisher Scientific #STEM00015) according to the manufacturer's instructions. For reprogramming cell lines, 500ng of hyPBase transposase were co-transfected with 750ng of PB-TAC-OKMS or PB-TAC-OKS vectors and 750ng of PB-TetO-AIO constructs using Lipofectamine™ Stem Transfection Reagent according to the manufacturer's instructions. Cells were then selected with 1µg/mL puromycin (Bio Basic #PJ593) for 4 days or 400µg/mL G-418 sulfate for 6 days.

Reprogramming towards iPSCs from mouse cells

Stably transfected Oct4-GFP ESC-derived NPC cell lines were seeded at 3 x 10^3 cells/cm^2 on a 0.1% porcine gelatin (Sigma-Aldrich #G1890) coating in NPC medium supplemented with 2% FBS (ThermoFisher Scientific #12483020) to promote adhesion. The following day, medium was replaced with FBSL supplemented with 1µg/mL doxycycline hyclate (Sigma-Aldrich #D9891) to induce overexpression of the reprogramming factors. Medium was refreshed every other day until D6. From D6 (unless mentioned otherwise), cells were maintained either in FBSL or 2i-LIF medium and passaged every 2-4 days depending on colonies density. For passaging, cells were treated with TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) and passaged as single-cells (1/3-1/30 ratio depending on density). mCherry+ and GFP+ cells were monitored using an Axio Vert.A1 (Zeiss) inverted fluorescence microscope with a 10X objective. Images were acquired with an Axiocam 305 color camera (Zeiss) and the ZEN 2 (blue edition) software (v3.9.100). Reprogramming cells were harvested using TrypLE™ Express Enzyme for dissociation, washed with DPBS and pelleted for RNA extraction or resuspended in 500µL of DPBS for analysis using a BD Accuri™ C6 Plus flow cytometer (BD Biosciences). iPSCs could readily be maintained in 2i-LIF medium after 20 days. All cell lines were negative for mycoplasma.

Assessment of mouse iPSC pluripotency in iPSC-embryo chimeras

Prior to aggregation, iPS cells were grown in KSR + 2i medium[116] consisting of high glucose DMEM with 15% KSR, 2mM GlutaMAX™, 1mM Na pyruvate, 0.1mM non-essential amino acids, 0.1mM 2-Mercaptoethanol (Gibco), 1000U/mL LIF, 5µg/mL Insulin (Sigma #I0516), 1µM PD0325901 (StemGent #04-0006), and 3µM CHIR-99021 (StemGent #04-0004).

For OKMS and OKS reprogramming: Embryos from superovulated CD-1 (Charles River) females were collected at 2.5 days post-coitum (dpc). Zonae pellucidae of embryos were removed using acid Tyrode's solution (Sigma) and clumps of cells from the iPS clones were obtained using Accutase (Millipore #SF006). Embryos were then aggregated with a clump of 6-10 iPSCs inside a depression well made in the dish with an aggregation needle (BLS Ltd). Aggregates were cultured overnight in microdrops of Life Global medium (Cooper Surgical LGGG-050) with 1mg/mL Protein Supplement (Cooper Surgical LGPS-020) covered by embryo-tested mineral oil (Cooper Surgical LGPO-500) at 37°C and 6% CO_2. Embryos were transferred the following morning into the uteri of 2.5-dpc pseudo pregnant CD-1 females. Chimerism was evaluated at birth based on the presence of black eyes and at weaning by coat color pigmentation. CD-1 outbred albino stock was used as embryo donors for aggregation with iPSCs and as pseudo pregnant recipients. Animals were maintained on 12h light/dark cycle and provided with food and water *ad libitum* in individually ventilated units (Techniplast) in the specific pathogen-free facility at The Centre for Phenogenomics (TCP). All procedures involving animals were performed in compliance with the Animals for Research Act of Ontario and the Guidelines of the Canadian Council on Animal Care. Animal Care Committee reviewed and approved all procedures conducted on animals at TCP.

For OKS + control and OKS + *Snhg26* reprogramming: An iPSC clamp consisting of 4-5 cells of a clone was aggregated with two tdTomato mouse embryos at the 8-cell morula stage in a depression well of a small drop of EmbryoMax® Advanced KSOM Embryo media (Sigma-Aldrich #MR-101-D) under mineral oil. Embryos had ubiquitous tdTomato expression derived from the Ai14 strain (RRID: IMSR_JAX:007914). After overnight incubation at 37°C with 5% CO2, the blastocysts were transferred to the uteri of pseudo pregnant females. The embryos were dissected at E8-8.5 stage. Contribution of

tdTomato negative iPSC-derived cells was visualized as negative cells/area in tdTomato-positive chimeric embryos. This was performed at the McGill Integrated Core for Animal Modeling (MICAM).

Assessment of mouse iPSC pluripotency using teratoma formation assay

Prior to the transplantation iPSCs were cultured as described above for aggregation. On the day of transplantation, iPSCs were harvested and dissociated into a single-cell suspension using Accutase (Millipore #SF006). The pellets were washed in PBS, and resuspended in Matrigel™ Basement Membrane (BD Biosciences #356234) - diluted 1 volume of Matrigel + 2 volumes of cold DMEM to get 5×10^6 cells per injection site. Mice were anesthetized using isoflurane, 100µL of cell suspension in diluted Matrigel were injected subcutaneously into each hind limb. Mice were monitored to determine when the tumor is palpable and visible, measured with calipers every 2-3 days to log tumor growth. Once the tumor has met end point guidelines, mice were euthanized and tumors dissected analysis. Briefly, teratoma were fixed in formalin and embedded in paraffin. Tissue cuts were stained by immunohistochemistry and scanned by Axioscan 7 (Zeiss) with a 20X objective. All procedures involving animals were performed in compliance with the Animals for Research Act of Ontario and the Guidelines of the Canadian Council on Animal Care. Animal Care Committee reviewed and approved all procedures conducted on animals at TCP.

Doubling time assays

ESC or NPC cell lines were seeded in replicates at 5×10^4 cells/cm^2 on a 0.1% porcine gelatin (Sigma-Aldrich #G1890) coating in 2i-LIF or NPC medium supplemented with 1% FBS (ThermoFisher Scientific #12483020) to promote adhesion. The following day, one replicate of cells was harvested using TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) and counted using a BD Accuri™ C6 Plus flow cytometer (BD Biosciences) as the starting timepoint (0h). Medium was replaced with 2i-LIF, NPC medium or FBSL medium supplemented with 1µg/mL doxycycline hyclate (Sigma-Aldrich #D9891) to induce transgenes overexpression. The cells were counted again after 24, 48, 72 and 120

hours. The total numbers of cells of 2-4 independent experiments were used to generate the average growth curves and to calculate the mean doubling times of the cell lines. Growth curves were represented with SEM using Microsoft Excel and doubling time graphs with STDs using GraphPad Prism (v7.04). The expression of an *mCherry* reporter gene allowed to calculate specifically the doubling time of the reprogramming cells. The cell specific growth rate (μ) was determined from the slope of the natural logarithm of cell count as a function of time and doubling time (DT) using the formula [DT = ln 2/μ]. Calculation used for doubling time was performed as follow: DT = (T2-T1) x ln(2)/ln(q2/q1) where T = time and q = quantity of cells. P-values between the different timepoints and doubling times were calculated using t-test (unpaired and unequal variance).

Limiting dilution assays

Stably transfected Oct4-GFP ESC-derived NPC cell lines were seeded at 5, 25, 125, or 500 cells per well of a Falcon™ 96-well Flat-Bottom microplate (Fisher Scientific #08-772-2C) on a 0.1% porcine gelatin (Sigma-Aldrich #G1890) coating in NPC medium. The following day, medium was replaced with FBSL medium supplemented with 1μg/mL doxycycline hyclate (Sigma-Aldrich #D9891) to induce the reprogramming into iPSCs. Medium was refreshed every other day for two weeks and switched at D18 to 2i-LIF medium until reprogramming completion. The number of wells negative for mCherry+ and GFP+ cells were quantified manually using an Axio Vert.A1 (Zeiss) inverted fluorescence microscope with a 10X objective after 12 and 27 days. In Microsoft Excel, the average percentage of negative wells was plotted on a semi-logarithmic graph in function of the number of cells seeded as an exponential line, with error bars representing the SEM. The Poisson distribution allowed to determine the number of cells ($x = 10\wedge(\frac{y-b}{a})$, where y = 37, a = slope and b = y intercept) that must be seeded to have 37% of negative wells. The frequency (1/x) and percentage (1/x *100) of cells able to reach the iPSC stage were calculated based on three independent experiments (n=3). Frequencies of reprogramming cells and iPSCs were represented

with STDs using GraphPad Prism (v7.04). P-values between the frequencies were calculated using t-test (unpaired and unequal variance).

Flow cytometry

Cells were harvested using TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) for dissociation, resuspended in 500µL of DPBS and filtered through a 35µm cell strainer (Fisher Scientific #08-771-23). Samples were loaded into a BD Accuri™ C6 Plus flow cytometer (BD Biosciences). Flow cytometry data were collected with the C6 Plus analysis software and analyzed with FlowJo™ (v10.8.1). Cell debris were gated apart using an FSC-A x SSC-A plot. Then, singlets were gated using an FSC-A x FSC-H plot and the gating for fluorescent reporter-positive cells was determined using a reporter-negative control sample as well as separated samples each positive for the different reporter genes. The total number of live single cells, as well as the percentages of mCherry+ and GFP+ cells and the volume used by the flow cytometer were collected directly from the .FCS files to calculate the density of mCherry+ and GFP+ cells as follow: $density\ of\ mCherry\ cells = \%\ of\ live\ single\ mCherry\ cells \times \left(\frac{number\ of\ live\ single\ cells}{cytometer\ uptake\ volume}\right)$. From this, the fraction of mCherry+ cells was calculated as follow: $\%\ mCherry\ cells = \frac{density\ of\ mCherry\ cells}{sum\ of\ mCherry, GFP\ and\ negative\ cell\ densities} \times 100$. The average values of 3-9 independent experiments were used and represented with the SEM using Microsoft Excel. P-values were calculated using t-test (paired samples).

ESC nuclear/cytoplasmic fractionation

Fractionation was performed following Gagnon *et al.*[117]. Briefly, R1 ESCs maintained in 2i-LIF medium on 0.1% porcine gelatin (Sigma-Aldrich #G1890) were expanded to a total of 3 x 10^6 cells. Cells were dissociated using TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021). 1 x 10^6 cells were kept for total RNA and the rest was processed for cell fractionation: Cells were resuspended in ice-cold RNAse-free hypotonic lysis buffer (10mM Tris pH7.5 (Bio Basic #TB0194), 10mM NaCl (Sigma-

Aldrich #S3014), 3mM MgCl2 (Sigma-Aldrich #M4880), 0.3% (vol/vol) IGEPAL® CA-630 (Sigma-Aldrich #I3021), 10% (vol/vol) glycerol (Fisher Scientific #BP229-1), 400U/mL of buffer of Murine RNase Inhibitor (NEB #M0314)) and incubated for 10 minutes on ice. The lysate was vortexed briefly and centrifuged 3 minutes at 4°C at 1,000g. The supernatant (cytoplasmic fraction) was collected on ice, added to 1mL of fresh RNA precipitation solution (5% 3M sodium acetate pH5.5 (Sigma-Aldrich #S8750) and 95% ethanol) and incubated for 1 hour at -20°C. The solution was vortexed briefly and centrifuged 15 minutes at 4°C at 18,000g before washing the pellet in ice-cold 70% (vol/vol) ethanol and centrifuging 5 minutes at 4°C at 18,000g prior to extraction. The pellet (nuclear fraction) was resuspended in ice-cold RNAse-free hypotonic lysis buffer and centrifuged 2 minutes at 4°C at 200g three times prior to extraction. Non-fractioned cells, cytoplasmic and nuclear fractions were finally resuspended in RLT buffer (Qiagen #74106) to extract RNA using RNeasy Kit (Qiagen #74106) according to the manufacturer's instructions. *Snhg26* lncRNA localization was evaluated by RT-qPCR by calculating the percentage of *Snhg26* cDNA PCR quantity in each fraction normalized to total *Snhg26* cDNA PCR quantity. *Xist, Malat1* and *Neat1* lncRNAs, and *Oct4* and *Klf2* mRNAs were used as controls for nuclear and cytoplasmic transcripts, respectively (qPCR primer sequences are listed in **Supplementary Table 8**). The average values of three independent experiments were used and represented with the STD using GraphPad Prism (v7.04). P-values were calculated using t-test (unpaired and unequal variance).

RNAScope *in situ* hybridization

R1 ESC colonies were grown in 2i-LIF medium on microscope slides coated with 0.1% porcine gelatin (Sigma-Aldrich #G1890). After 3 days, ESCs were fixed with 4% paraformaldehyde (Fisher Scientific #AA433689L) for 30 minutes at room temperature, sequentially dehydrated using 50%, 70% and 100% ethanol for 5 minutes at room temperature and stored at -20°C. *In situ* hybridization of a set of probes targeting *Snhg26* transcript (Advanced Cell Diagnostics #1574981) was performed using the RNAScope® Multiplex Fluorescent Reagent Kit v2 (Advanced Cell Diagnostics #323100) according to the manufacturer's instructions. Briefly, slides were sequentially re-hydrated using 100%, 70% and 50%

ethanol for 5 minutes at room temperature. Cells were then permeabilized using 0.1% Tween 20 (Fisher Scientific #BP337-500). Slides were incubated for 10 minutes at room temperature after addition of H2O2. Protease III digestion was performed for 10 minutes at room temperature. The *Snhg26* probes were hybridized for 2 hours at 40°C. RNAScope® amplification steps with preamplifiers AMP1 (30 minutes), background reducers AMP2 (30 minutes) and amplifiers AMP3 (15 minutes) were then performed at 40°C. Amplifiers were conjugated to 1:2000 TSA Vivid Fluorophore 650 (Advanced Cell Diagnostics #323273). Nuclear staining was performed using DAPI from the RNAScope® Kit and slides were mounted in ProLong™ Gold Antifade Mountant (ThermoFisher Scientific #P36930). Imaging was performed with an Axio ObserverZ.1 (Zeiss) microscope with LSM900 confocal head using 3µm z-steps. LD C-Apochromat 40X 1.1 water objective was used and images were acquired with the ZEN (blue edition) software (v3.3). The following parameters were used: LSM scan speed 4; DAPI 500V detector gain and 3.5% laser wavelength; Cy5 (TSA Vivid Fluorophore 650) 500V detector gain and 0.2% laser wavelength. Stacks were reconstructed using the Z project plugin in Fiji (ImageJ v2.14.0 Fiji).

Knock-down of *Snhg26* in mouse and human ESCs

The inducible knock-down cell lines were generated using CRISPRi-based methods following Gilbert *et al*.[91]. Single guide RNAs (sgRNAs) were designed using the CRISPR 10K Tracks of the UCSC genome browser[118] and ordered from IDT™. They were resuspended and annealed before their ligation into an in-house PB-U6-Gateaway plasmid digested beforehand with BbsI-HF® (NEB #R3539) using a T4 DNA ligase (ThermoFisher Scientific #EL0016) according to the manufacturer's instructions (sgRNA sequences are listed in **Supplementary Table 8**). The sequences of all plasmids were verified by Sanger sequencing. Complete sequences of all plasmids are available upon request.

For mouse *Snhg26* knock-down, 400ng of hyPBase transposase were co-transfected with 850ng of PB-TetO-AIO-dCas9-BFP-KRAB vector and 350ng of control or *Snhg26*-targeting sgRNA constructs using Lipofectamine™ Stem Transfection Reagent according to the manufacturer's instructions. Cells were then selected with 1µg/mL puromycin (Bio Basic #PJ593) for 4 days or 400µg/mL G-418 sulfate for 6

days. The mouse ESC lines were seeded prior to the knock-down at 2.6 x 10^4 cells/cm^2 on 0.1% porcine gelatin (Sigma-Aldrich #G1890) in 2i-LIF medium. The knock-down was induced by adding 1 µg/mL doxycycline hyclate (Sigma-Aldrich #D9891) and ESCs were harvested after 5 days for RNA extraction. For human *SNHG26* knock-down, primed hESC lines were seeded at a density of 2.6 x 10^4 cells/cm^2 on Geltrex™-coated plates (hESC-qualified, growth factor-reduced, ThermoFisher Scientific, #A1413302) in mTeSR™1 medium (StemCell Technologies, #85850) supplemented with 10µM Y-27632 (Tocris Bioscience #1254). The following day, medium was replaced by fresh mTeSR™1 medium and cells were co-transfected with 500ng of hyPBase transposase, 1µg of PB-TetO-AIO-dCas9-BFP-KRAB vector and 875ng of control or four *SNHG26*-specific pooled sgRNAs using Lipofectamine™ Stem Transfection Reagent (ThermoFisher Scientific #STEM00015) according to the manufacturer's instructions. Cells were then selected with 500ng/mL puromycin (Bio Basic #PJ593) for 6 days. The hESCs were seeded prior to the knock-down on Geltrex™-coated plates (hESC-qualified, growth factor-reduced, ThermoFisher Scientific, #A1413302) in mTeSR™1 medium (StemCell Technologies, #85850). The knock-down was induced by adding 1 µg/mL doxycycline hyclate and ESCs were harvested after 2 or 3 days for RNA extraction.

Snhg26 and *SNHG26* knock-down efficiencies as well as the mean relative fold-change in pluripotency-related mRNAs expression were measured via RT-qPCR by comparing to the control condition (non-targeting sgRNA) (qPCR primer sequences are listed in **Supplementary Table 8**). The average values of 2-3 independent experiments were used and represented with the STD using GraphPad Prism (v7.04). P-values were calculated using t-test (unpaired and unequal variance).

For analysis of global gene expression through lrRNAseq: The gene expression values in transcripts per million reads (TPM) of hESCs with control or *SNHG26* sgRNAs at D2 or D3 were averaged (**Supplementary Table 8**). Significantly up- or down-regulated genes were established either as genes with a fold change ≥ 1.5 or ≤ 0.67, respectively, and adjusted p-value ≤ 0.05 (any reference to adjusted p-values throughout the text is done using the Benjamini-Hochberg (FDR) method); or as genes with a Shannon entropy score below tenth percentile cutoff of the entropy scores distribution as previously

described[16]. The Shannon entropy modelling to compute an expression-specificity index for each gene was generated as previously described[16,119]. Briefly, for each gene, a relative expression (Ri) value was calculated per timepoint (i.e. D2 or D3) and per condition (i.e. control or *SNHG26* sgRNAs), where Ri = (expression per timepoint or average expression per condition)/sum of TPM values in all the timepoints or all the conditions. The entropy index score (Hi) across all samples or groups was calculated as $Hi = -1 \times \Sigma(Ri \times \log2(Ri))$. Represented down-regulated genes have a significance based either on fold change and adjusted p-value or on Shannon entropy score. Heat map expression of significant genes (n=1,794) was generated using the TPM expression values of these genes using GraphPad Prism (v7.04), setting the maximum color for values ≥ 15 TPM. Represented up- and down-regulated genes have a significance based only on Shannon entropy score (**Supplementary Table 8**). GO-term enrichment was assessed using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[120] (v2022q3) on the genes significantly up- or down-regulated based on the Shannon entropy score only. Significant GO-terms were selected based on a fold enrichment ≥ 1.5, a p-value ≤ 0.05, and an FDR ≤ 0.1 (**Supplementary Table 8**).

Reprogramming of human fibroblasts towards iPSCs

Human reprogramming was performed using mRNAs generated by the HiScribe® T7 mRNA Kit with CleanCap® Reagent AG (NEB #E2080S) using constructs for *OCT4, SOX2, KLF4, C-MYC, LIN28* and nuclear-destabilized *GFP* (NDG) described in Mandal *et al*.[121]. BJ fibroblasts (ATCC #CRL-2522) were plated at 20,000 cells per well of a 24-well plate and transfected with 150ng of mRNA cocktail containing a 3:1:1:1:1:1 molar ratio for *OCT4, SOX2, KLF4, C-MYC, LIN28*, and *NDG* for 10 days with Lipofectamine™ RNAiMAX (ThermoFisher Scientific #13778030). For the first 12 days, cells were cultured in NutriStem® hPSC XF GF-free (Sartorius #06-5100-01-1A) that was conditioned on mitotically arrested human neonatal dermal fibroblasts (amsbio #DFN-F) treated with mitomycin C for 3 hours. Starting at D13, cells were cultured in in-house generated MEF-conditioned media on mitotically-

arrested mouse embryonic fibroblasts treated with mitomycin C for 3 hours and generated by the Lunenfeld-Tanenbaum Research Institute.

Knock-down of *SNHG26* in human reprogramming

Knock-down of *SNHG26* during human reprogramming began with transfection of parallel, antisense oligonucleotides (ASOs) control or targeting *SNHG26* lncRNA (ASOs were custom-ordered and generated by IDT™, sequences are listed in **Supplementary Table 8**) at D3 using Lipofectamine™ RNAiMAX (ThermoFisher Scientific #13778030) with 30nM of ASO. Knock-down was subsequently performed every 3 days. Cells were passaged 1/3 at D13 and fixed at D20 before staining by immunofluorescence with NANOG (R&D Systems #AF1997), OCT4 (Santa Cruz Biotechnology #sc-9081), and SOX2 (R&D Systems #MAB2018) antibodies. Nuclear staining was performed using DAPI (Sigma #D9542-1MG) and whole-well immunofluorescence imaging was performed with the InCell Auto 3000 (GE Healthcare). Quantification of NANOG, OCT4, and SOX2-positive colonies was performed manually. Quantification of OCT4-positive colonies area was performed in Fiji (ImageJ v1.54p Fiji) by smoothing granularity with a *Gaussian blur* function at sigma = 4µm. Thresholding was performed by adjusting the upper slider to 40. Threshold value was visually validated to ensure complete segmentation of OCT4-positive colonies while minimizing background inclusion. Noise was removed by using the *open* function once, touching colonies were then separated by the *watershed* function, and *particle size* was set > 5,000µm^2 to automatically count the regions of interest. The area values of 175 to 1,042 colonies were averaged and represented with the SEM using Microsoft Excel. P-values were calculated using t-test (unpaired unequal variance). *SNHG26* knock-down efficiency was measured at D12 via purifying the RNA using Qiagen RNeasy Kit (Qiagen #74106) and reverse-transcription using Maxima H-minus Reverse (ThermoFisher Scientific #EP0751), followed by qPCR with the LightCycler 480 SYBR Green kit (Roche #04707516001), by comparing to the control condition (control ASOs) (qPCR primer sequences are listed in **Supplementary Table 8**).

1 ***Snhg26* knock-out (KO) in mouse ESCs**

2 *Snhg26*^{KO} R1 ESC lines were established as described in Ran *et al*.[122], using two sgRNAs to remove the
3 full fourth exon of *Snhg26*. SgRNAs were designed using the CRISPR 10K Tracks of the UCSC genome
4 browser[118] and ordered from IDT™. 0.350ng of annealed sgRNA pairs were cloned in 50ng of the pCCC
5 backbone[123] (gifted by Dr. Francis Lynn, University of British Columbia) by BbsI HF® digestion (NEB
6 #R3539) followed by ligation using a T4 DNA ligase (ThermoFisher Scientific #EL0016) according to
7 the manufacturer's instructions (sgRNA sequences are listed in **Supplementary Table 8**). The sequences
8 of all entry clone inserts were verified by Sanger sequencing. Complete sequences of all plasmids are
9 available upon request.

10 For targeting with the CRISPR-Cas9 system, ESC lines were seeded at a density of 13 x 10^3 cells/cm^2 on
11 plates coated with 0.1% porcine gelatin (Sigma-Aldrich #G1890) in 2i-LIF medium. The following day,
12 cells were transfected with 100ng of hyPBase transposase, 200ng of transposon construct corresponding
13 to either the control cDNA of *Luciferase-antisense*, *Snhg26.217* or *Snhg26.216*, and 800ng of each
14 sgRNA construct using Lipofectamine™ Stem Transfection Reagent (ThermoFisher Scientific
15 #STEM00015) according to the manufacturer's instructions. Cells were then selected with 1µg/mL
16 puromycin (Bio Basic #PJ593) for 4 days. Five days post-selection, cells were dissociated as single-cells,
17 counted and seeded at one cell per well of a Falcon™ 96-well Flat-Bottom Microplate (Fisher Scientific
18 #08-772-2C) with a multichannel pipettor for colony isolation. Medium was replaced every other day and
19 wells containing single colonies were selected by eye using an Axio Vert.A1 (Zeiss) inverted microscope
20 with a 10X objective. One week post-seeding, clones were split, expanded and frozen prior to screening
21 and further characterization. All cell lines were maintained in 2i-LIF medium on 0.1% porcine gelatin, in
22 a humidified incubator at 37°C under 5% CO2 and were negative for mycoplasma.

23 For *Snhg26*^{KO} ESC clone screening and validation, genomic DNA (gDNA) was extracted using gSYNC™
24 DNA Extraction Kit (Geneaid #GS100) following the manufacturer's instructions, with 100µg/mL
25 RNaseA treatment (ThermoFisher Scientific #EN0531) 20 minutes at room temperature. DNA was

quantified using a NanoDrop™ 2000 Spectrophotometer (ThermoFisher #ND-2000C) and 300ng of DNA were loaded onto a 1% Agarose A (Bio Basic #D0012) gel to verify the DNA integrity and purity. Assessment of ESC clones genotype regarding *Snhg26* locus was performed by PCR on 2ng of gDNA using the dNTP Solution Set (NEB #N0446S) and the Taq DNA polymerase (NEB #M0267) according to the manufacturer's instructions. 15-25µL of PCR-amplified product was loaded on a 1% agarose gel with 5µL of the GD 1kb plus DNA Ladder RTU (FroggaBio #DM015-R500F) and revealed using an Azure 200 gel image (Azure Biosystems). PCR-amplification of *Snhg26* fourth exon screening followed by Sanger sequencing validation was used to assess the copy number of *Snhg26* allele in each clone (genotyping primer sequences are described in **Supplementary Table 8**).

Blastoid differentiation and characterization

The blastoids were generated following a protocol adapted from Jana *et al.*[124]. Briefly, wild-type or *Snhg26*^{KO} mouse ESC lines routinely cultured in 2i-LIF medium were seeded at 9.4 x 10^3 cells/cm^2 on a 0.1% porcine gelatin (Sigma-Aldrich #G1890) coating in serum + LIF (SL) medium and passaged three times using TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021). On the third passage, SL medium was supplemented with 1µM PD0325901 (Selleck Chemicals #S1036) for 48 hours. Then, an Aggrewell™ 400 plate (StemCell Technologies #34411) was treated with Anti-Adherence Rinsing Solution (StemCell Technologies #07010) as per the manufacturer's instructions, and 48,000 cells were seeded per well in RBF medium. RBF media was composed of 12.5g/L GMEM with 10% FBS (vol/vol), 32.7mM NaHCO3, 1mM sodium pyruvate, 0.1mM non-essential amino acids, 0.1mM β-Mercaptoethanol, supplemented with 1µM retinoic acid, 25ng/mL bFGF and 20ng/mL BMP4. 24 hours later, 500µL of RBF medium was carefully added to each well. On the following day, 500µL of RBF medium containing CHIR-99021 (Selleck Chemicals #S2924) was added to each well for a final concentration of 3µM.

72h post seeding, brightfield images of the blastoids were acquired using an Axio Vert.A1 (Zeiss) microscope with a 5X objective coupled to an Axiocam 305 color camera (Zeiss) and the ZEN 2 (blue

edition) software (v3.9.100). Blastoids were then carefully transferred from into two 1.5mL Eppendorf® LoBind® microcentrifuge tubes (Avantor #CA80077-230) using ART™ wide bore filtered pipette tips (ThermoFisher Scientific #2079G). They were washed with DPBS (ThermoFisher Scientific #14190250) once by centrifuging 2 minutes at 300g at room temperature. Finally, blastoids were fixed with 4% paraformaldehyde (methanol-free, Fisher Scientific #AA433689L) for 30 minutes at 4°C under light agitation and rinsed in DPBS three times. Permeabilization was performed in DPBS containing 0.1% Tween (Fisher Scientific #BP337-500) (DPBS-T 0.1%) for 10 minutes and blastoids were blocked for 1 hour in 5% donkey serum (Sigma-Aldrich #D9663) diluted in the DPBS-T 0.1%. Primary antibodies against SOX2 (Cell Signaling Technology #23064) and GATA6 (Novus Biologicals #AF1700) were respectively diluted 1:100 and 1:80 in the blocking solution and incubated at 4°C overnight under light agitation. Blastoids were then rinsed three times in DPBS-T 0.1% for 5 minutes and incubated with secondary antibodies Alexa™ Fluor 488 anti-goat (ThermoFisher Scientific #A21467) and 594 anti-rabbit (ThermoFisher Scientific #A21207) diluted 1:500 in the blocking solution for 60 minutes at RT. Following 3 rinsings with DPBS-T 0.1%, samples were incubated with DAPI (Sigma-Aldrich #D9542) diluted 1:200 for 7 minutes at RT, before 3 rinsings in DPBS and mounting on Fisherbrand™ Superfrost™ Plus microscope slides (Fisher Scientific #12-550-15) on which two SecureSeal™ Imaging Spacers (Sigma-Aldrich #GBL654002-100EA) were placed and filled each with 10µL containing the stained blastoids in ProLong™ Gold Antifade Mountant (ThermoFisher Scientific #P36930). Imaging was performed with an Axio ObserverZ.1(Zeiss) microscope with LSM900 confocal head using 2µm z-steps. LD C-Apochromat 40X 1.1 water objective was used and images were acquired with the ZEN (blue edition) software (v3.3). The following parameters were used: LSM scan speed 6; DAPI 575V detector gain and 3% laser wavelength; 488 600V detector gain and 3% laser wavelength; 594 540V detector gain and 2.8% laser wavelength. Stacks were reconstructed using the Z project plugin in Fiji (ImageJ v2.14.0 Fiji). Channels on the reconstructed Z images were split and analyzed separately. In Fiji, the *Freehand* tool was used to manually draw the outlines of each blastoid and measure the area. The total number of cells per blastoids was quantified manually based on DAPI-positive cells. SOX2- and GATA6-

positive cells were counted manually. The average values of area, total number of cells, SOX2- or GATA6-positive cells in the wild-type (n=9-16), *Snhg26*^{KO} + control (n=9-12), *Snhg26*^{KO} + *Snhg26.217* (n=10-12), or *Snhg26*^{KO} + *Snhg26.216* (n=10-14) blastoids were used and represented as box plots with the minimal and maximal values using GraphPad Prism (v7.04). P-values were calculated by ordinary one-way ANOVA.

RNA extraction and reverse transcription-qPCR

Cell lysates were homogenized with QIAshredder columns (Qiagen #79656) and total RNA was extracted using RNeasy Kit (Qiagen #74106) according to the manufacturer's instructions. RNA was quantified using a NanoDrop™ 2000 Spectrophotometer (ThermoFisher #ND-2000C) and then DNase-treated using DNaseI (ThermoFisher Scientific #EN0521) according to the manufacturer's instructions. 300ng of DNase-treated and non-treated RNA were loaded onto a 1% Agarose A (Bio Basic #D0012) gel to verify the RNA integrity and purity. cDNA was synthesized from 500ng of RNA by oligo-dT and random priming with iScript™ gDNA Clear cDNA Synthesis Kit (Bio-Rad #1725035), according to the manufacturer's instructions. cDNA was diluted ten times and qPCRs were performed using SYBR™ Select Master Mix (ThermoFisher Scientific #4472908) on a CFX384 Touch real-time PCR detection system (Bio-Rad). Raw data was retrieved using the Bio-Rad CFX Maestro (v5.3.022.1030) software. Relative gene quantification was analyzed based on the $2^{-\Delta\Delta Ct}$ method by applying the comparative cycle threshold (Ct) method on Microsoft Excel. Relative expression levels were normalized to *Rpl27 or Rps13* and to *RPS13* housekeeping genes for mouse and human cells analyses, respectively (qPCR primer sequences are listed in **Supplementary Table 8**).

Library preparation and lrRNAseq, processing and alignment

Cell lysates were homogenized with QIAshredder columns (Qiagen #79656) and total RNA was extracted using RNeasy Kit (Qiagen #74106) according to the manufacturer's instructions (RNAseq samples are listed in **Supplementary Table 1**). RNA was quantified using a NanoDrop™ 2000 Spectrophotometer

(ThermoFisher #ND-2000C) and then DNase-treated using TURBO™ DNase (ThermoFisher Scientific #AM2238) according to the manufacturer's instructions. 300ng of DNase-treated and non-treated RNA were loaded onto a 1% Agarose A (Bio Basic #D0012) gel to verify the RNA integrity and purity. RNA was also quantified using the Qubit™ RNA Broad Range Kit (ThermoFisher Scientific #Q10210) on Qubit™ 4 Fluorometer (ThermoFisher Scientific #Q33238) according to the manufacturer's instructions. RNA quality was further verified using an RNA 6000 Nano Kit (Agilent Technologies #5067-1511) on a 2100 Bioanalyzer instrument (Agilent Technologies #G2939B) according to the manufacturer's instructions. Samples all had RNA integrity numbers of 7.8 or more.

For samples sequenced on R9 Flow Cells, polyadenylated RNA was enriched from 7.5-20µg of total RNA using NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB #E7490) according to the manufacturer's instructions. Poly(A)+-enriched RNA was quantified using the Qubit™ RNA High Sensitivity Kit (ThermoFisher Scientific #Q32852) on Qubit™ 4 Fluorometer according to the manufacturer's instructions and poly(A)+-enriched RNA smear quality was verified using an RNA 6000 Pico Kit (Agilent Technologies #5067-1513) on a 2100 Bioanalyzer instrument according to the manufacturer's instructions. Libraries were generated using 55-100ng of poly(A)+ RNA and a Direct cDNA Sequencing Kit (Oxford Nanopore Technologies (ONT) #SQK-DCS109), with each sample barcoded with the Native Barcoding Expansion (ONT #EXP-NBD104), according to the manufacturer's protocol: "Direct cDNA Native Barcoding (SQK-DCS109 with EXP-NBD104 and EXP-NBD114)" version DCB_9091_v109_revF_04Feb2019. Barcoded samples were pooled, loaded on a MinION sequencer (ONT) using eight R9.4.1 Flow Cells (ONT #FLO-MIN106D), and sequenced using the MinKNOW™ (v19) software. Purification of DNA between each step was performed with KAPA HyperPure Beads (Roche #08963835001). R9.4.1 Flow Cells were washed and reloaded with the Flow Cell Wash Kit (ONT #EXP-WSH003) to increase sequencing depth obtained per Flow Cell.

For samples sequenced on R10 Flow Cells, Ligation sequencing amplicons - Native Barcoding Kit 24 V14 (ONT #SQK-NBD114.24) was used with 50ng of total RNA. Each sample was assigned a specific barcode. Briefly, an oligo-dT VN Primer (VNP, sequence: 5'-

5phos/ACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTTTVN-3', where V=A, C, or G, and N=A, C, G, or T) was used to reverse-transcribe poly(A) RNA into cDNA with the Maxima H Minus Reverse Transcriptase (RT) (ThermoFisher Scientific #EP0751). A Strand Switching Primer (SSP, sequence: 5'-TTTCTGTTGGTGCTGATATTGCTmGmGmG-3', where mG = 2' O-Methyl RNA bases) was used for strand switching during the RT reaction. Libraries were PCR-amplified for 14 PCR cycles with ONT's cDNA Primers (cPRM) (Forward primer sequence: 5'-ATCGCCTACCGTGACAAGAAAGTTGTCGGTGTCTTTGTGACTTGCCTGTCGCTCTATCTTC-3', Reverse primer sequence: 5'-ATCGCCTACCGTGACAAGAAAGTTGTCGGTGTCTTTGTGTTTCTGTTGGTGCTGATATTGC-3') and the LongAmp® Taq 2X Master Mix (NEB #M0287) or the NEBNext Ultra II Q5 Master Mix (NEB #M0544L), following ONT's SQK-PCS109 kit recommendations for PCR incubation times and temperatures. cDNA was quantified using the Qubit™ 1X dsDNA High Sensitivity Kit (ThermoFisher Scientific #Q33230) on Qubit™ 4 Fluorometer according to the manufacturer's instructions and cDNA smear quality was verified using a DNA 12000 Kit (Agilent Technologies #5067-1508) on a 2100 Bioanalyzer instrument (Agilent Technologies #G2939B) according to the manufacturer's instructions. The blunt DNA ends were prepared using the NEBNext® Ultra™ II End Repair/dA-Tailing module (NEB #E7546), adding a 3' dA tail and phosphorylating the 5' end. Barcodes were ligated to the DNA fragments with a Blunt/TA ligase (NEB #M0367), barcoded samples were pooled and motor protein adapters were finally ligated with the NEBNext® Quick Ligation Module (NEB #E6056). Purification of DNA between each step was performed with KAPA HyperPure Beads (Roche #08963835001). Library was loaded on a PromethION 2 Solo (ONT) sequencer using R10.4.1 Flow Cells (ONT #FLO-PRO114M) and sequenced using the MinKNOW™ (v23.07.15) software.

Raw fast5 files from sequencing with R9.4.1 Flow Cells were converted to the more recent pod5 format using the pod5-file-format software from ONT's "nanoporetech" GitHub page (https://github.com/nanoporetech/pod5-file-format). pod5 files were then basecalled using Guppy (v6.5.7) using super accurate settings with the guppy_basecaller command and the dna_r9.4.1_450bps_sup.cfg

configuration file. Barcodes were detected and reads were demultiplexed with the guppy_barcoder command using the –barcode_kits "EXP-NBD104 EXP-NBD114" argument. Raw pod5 files from sequencing with R10.4.1 Flow Cells were basecalled with Guppy (v6.5.7) using super accurate settings with the guppy_basecaller command and the dna_r10.4.1_e8.2_400bps_sup.cfg configuration file. Barcodes were detected and reads were demultiplexed with the guppy_barcoder command using the –barcode_kits "SQK-NBD114-24" argument. The resulting fastq reads were aligned to the GRCm38/mm10 mouse and GRCh38/hg38 human reference genomes with Minimap2[125] (v2.24-r1122) using the following parameters: -aLx splice –cs = long. Aligned reads were quality-controlled with the NanoPlot software[126] (v1.43.2).

Mouse transcriptome assembly and isoform classification

To properly assemble the transcriptome file, reads from each sample alignment file were oriented according to transcript's proper strand relative to the genome using alignment files obtained from Minimap2. To do so, we used the following strategy per sample file:

1) Reads were first filtered to keep reads with MAPQ > 13.
2) Multi-exonic/split reads (i.e. reads with splicing information) were re-oriented to reflect gene orientation with respect to the genomic strand. Minimap2 aligned and contain a "ts" tag to reflect transcript strandedness or sequence orientation (i.e. sense (+) versus antisense (-) sequence) but not the genomic strand. So here, we divided these split reads into 4 categories:
    a. Reads with a 'ts:A:+' tag, which represents sense transcript sequence, and a forward read orientation with respect to the genome (i.e. SAM flag = 0) were isolated using piped "samtools view -F 16" (Samtools v1.17) and "grep -E 'ts:A:\+'" functions and maintained in the same read orientation. They were given an 'XS:A:+' tag designation by modifying the 'ts:A:+' tag field using the "sed" function within the SAM alignment files. This is done to reflect a forward gene and read orientation with respect to the genome.

b. Reads with an ‘ts:A:+’ tag and a reverse read orientation with respect to the genome (i.e. SAM flag = 16) were isolated using piped “samtools view -f 16” and “grep -E ‘ts:A:\+’” functions and maintained in the same read orientation. However, they were given an ‘XS:A:-‘ instead of the ‘ts:A:+’ tag to switch the gene orientation to reflect antisense or reverse gene orientation with respect to the genome.

c. Reads with a ‘ts:A:-’ tag, which represent antisense transcript sequence, and a forward read orientation with respect to the genome (i.e. SAM flag = 0), were isolated using piped “samtools view -F 16” and “grep -E ‘ts:A:-’” functions and re-oriented by changing the SAM flag to 16 (i.e. reverse read orientation) using the “sed” function in terminal and were given an ‘XS:A:-‘ designation as done above, to reflect both a reverse gene and read orientation with respect to the genome.

d. Reads with a ‘ts:A:-’ tag and a reverse read orientation with respect to the genome (i.e. SAM flag = 16), were isolated using piped “samtools view -f 16” and “grep -E ‘ts:A:-’” functions and re-oriented by changing the SAM flag to 0 (i.e. forward read orientation) using the “sed” function in terminal. However, they were given an ‘XS:A:+‘ designation as done above, to switch the gene orientation to reflect sense gene orientation with respect to the genome.

3) For mono-exonic transcripts, the reads were oriented with the Pychopper software from ONT's “epi2me-labs” github page (https://github.com/epi2me-labs/pychopper).

4) All reads from the above filtered and re-oriented files were merged using Samtools to generate one alignment file per sample.

The transcriptome assembly was performed on each sample BAM file using the TAMA software (tc_version_date_2022_02_22)[127], with the tama_collapse.py script and “-x no_cap -c 60 -i 75 -icm ident_map -sj sj_priority” parameters. Transcriptomes derived from each sample were then merged using the tama_merge.py script to produce one assembled transcriptome BED file for the whole dataset. To obtain GTF files, we used the tama_convert_bed_gtf_ensembl_no_cds.py script to convert TAMA

generated BED files to GTF files. The resulting GTF files were then processed with the isoform classification pipeline of the SQANTI3 software[54] (v5.1.1) using the sqanti3_qc.py script and the "--CAGE_peak" and "--polyA_motif_list" options to identify transcription start and end sites, respectively, to classify isoforms using GRCm38/mm10 as reference genome and Ensembl's gene annotations version 99 as reference transcriptome. The --isoAnnotLite argument was used with Ensembl's annotations version 86 to generate a GFF3 file containing functional and structural annotations for novel isoforms (available as **Supplementary Data 4**). Isoforms that are considered artifacts were then filtered using SQANTI3 filter with the "Rules filter" using default rules and sqanti3_filter.py script. Filtered-out reference isoforms were then rescued using SQANTI3 sqanti3_rescue.py script and the "rules" argument.

The SQANTI3 produced classification file included the structural categories and subcategories, predictions for RT switching, coding potential and non-sense mediated decay (NMD), and detection of transcription start sites (TSS) and poly(A) motifs. A total of 1,883,390 were identified. However, isoforms with RT switching events and without poly(A) motif were removed to reach a total of 1,038,414 isoforms. Then, IsoQuant[51](v3.9.0) was used to quantify the 1,038,414 isoforms, and low expression isoforms (isoforms where expression is ≥ 1 TPM in at least 3 samples) were filtered out to obtain 288,098 transcripts. IsoQuant was re-used to re-quantify and filter out lowly expressed transcripts from this smaller list of 288,098 transcripts and more accurately quantify and obtain a final list of 279,998 isoforms (see **Supplementary Fig. 2** and **Supplementary Table 2**). The coding potential score of detected transcripts was assessed with the Coding-Potential Assessment Tool (CPAT)[128] (v3.0.5) (**Supplementary Table 2**), which accounts for RNA sequence characteristics such as the ORF size and coverage, using default parameters with the exception of "--min-orf", which was set at 10. P-values between the distributions of coding probability were calculated by t-test (unpaired and unequal variance).

Analysis of publicly available bulk and single-cell short-read RNAseq

Raw 5' RNAseq fastq files from O'Malley *et al.*[129] were downloaded from Array express E-MTAB-1654. Raw paired-ends RNAseq fastq files from Stadhouders *et al.*[72] were downloaded from GEO accession

GSE96608. Raw paired-ends RNAseq fastq reads from Qiao *et al.*[13] were downloaded from GEO accession number GSE138760. Raw single-cell RNAseq fastq files from Kim *et al.*[18] were downloaded from GEO accession number GSE55291. Read quality was assessed with fastQC (v0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were aligned to the mouse mm10/GRCm38 genome using the STAR aligner[130] (v2.7.11a) with the following parameters: --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within --outSAMstrandField intronMotif --outSAMattributes Standard . Resulting BAM files were indexed and alignment statistics exported with Samtools (v1.17).

For the O'Malley *et al.* dataset, every reprogramming sample was merged using Samtools merge. Coverage for 5' nucleotides across each read in every sample was calculated with bedtools[131] (v2.30.0) using genomecov -5 -bg . Both strands were processed separately using the strand + and strand - arguments respectively. 5' coverages were merged as peaks and coverage was summed per peak with the bedtools merge command using the following parameters: -d 10 -c 4 -o sum. Bed files for plus and minus strand peaks were concatenated and sorted by position. Coverage was converted to counts per million (CPM), and peaks with a coverage below 0.5 CPM were filtered out. The resulting bed file was used in addition to CAGEseq data for SQANTI3 analysis, presented in the **Preparation of a catalog of validated isoform 5' ends** section.

For the Stadhouders *et al.* and Qiao *et al.* datasets, Salmon[132] (v1.10.3) was used with the mapping-based mode to quantify expression of a subset of isoforms of interest (available in **Supplementary Table 6**). First, an index was generated using the salmon index command with a fasta file of the isoforms as transcripts and the mouse mm10 chromosomes as decoys. The salmon quant command was used on fastq format reads with the previously generated index and the following arguments: -l U –validateMappings. For Stadhouders *et al.*, TPMs were averaged across reprogramming stages based on principal component analyses presented in Stadhouders *et al*: B cells and Bα as "B cells", D2 and D4 as "early", D6 and D8 as "late". TPM values are available in **Supplementary Table 6** and **Supplementary Table 8**.

For the Kim *et al*. single-cell RNAseq dataset, the salmon quant command was used on fastq format reads with the previously generated index (see the paragraph above) and the following arguments: -l U –validateMappings. To differentiate high-expressor cells from low-expressors and background noise, cells were considered positive for expression of a given isoform if its expression exceeded a per-isoform threshold set at the 33rd percentile. Gene-level read counts were obtained with the featureCounts tool (v2.0.3), from the Subread package, using the exon counting mode[133]. Resulting counts matrices were processed in R (v4.3.3) using Seurat (v5.3.1). Data was processed using Seurat's recommended SCTransform workflow, as outlined in the official Seurat vignettes. Module scores showing normalized expression of gene sets were calculated using Seurat's AddModuleScore, using 100 control genes per analyzed gene. Significant differences in module scores were assessed with a Mann-Whitney non-parametric U-test. Log-normalized counts, isoform level TPM values, gene sets and module scores are available as **Supplementary Table 6**.

Analysis of publicly available ChIPseq data

H3K4me3 ChIPseq and input data from Hussein *et al*.[16] was re-analyzed. Fastq files were first trimmed with trim_galore (v0.6.10) (https://github.com/FelixKrueger/TrimGalore) using the --paired argument and quality was evaluated with fastQC. Reads were aligned to the NCBI38/mm10 mouse reference using the Bowtie2 alignment algorithm[134] (v2.5.4) with the pre-set sensitive parameter. Peaks were called using MACS3 (v3.0.3) callpeak using the following arguments: -g mm -f BAMPE -nomodel -B . Resulting bed files from each reprogramming sample were concatenated and sorted by position, then overlapping peaks were merged with bedtools merge.

LrRNAseq data analysis of gene expression

Raw read counts were obtained using IsoQuant[51] (v3.9.0) using default parameters for gene level counts. Together, the sequencing runs (45 samples in total; **Supplementary Table 1**) generated 94,777,253 quality control-passed and filtered reads in total, with an average of 2,106,161 reads and 3.8 Gb of data

per replicate (**Supplementary Table 1**). EdgeR R-package[135] (v3.42.4) was then used to normalize the data, calculate RNA abundance at the gene level (as CPM), and perform statistical analysis. Briefly, a common biological coefficient of variation (BCV) and dispersion (variance) were estimated based on a negative binomial distribution model. This estimated dispersion value was incorporated into the final EdgeR analysis for differential gene expression, and the generalized linear model (GLM) likelihood ratio test was usegd for statistics, as described in EdgeR user guide. Genes with CPM ≥ 1 in at least one sample were considered (n=21,757 genes) and listed in **Supplementary Table 1**. Differential gene expression (DEGs) was established as genes with a fold change ≥ 2 (up-regulation) or ≤ 0.5 (down-regulation) and adjusted p-value (FDR) ≤ 0.05. Gene expression plots were generated using Microsoft Excel by averaging the log10(CPM+1 expression values) of the different replicates of a given timepoint for the different gene lists and representing the SEM as a shaded contour (gene lists are listed in **Supplementary Table 1** and **Supplementary Table 4**). Heat map expression of manually curated alternative splicing regulators was generated using the minimum and maximum log10(CPM+1) expression values of these genes on Microsoft Excel (**Supplementary Table 4**).

Sample clustering was performed on top varying genes: Briefly, the variance of expression was calculated for each gene. All genes with a variance across samples superior to 10% above the average variance were considered as top-varying genes (n=489 genes). Gene expression was normalized as z-scores, and then distance matrices were calculated with the *dist* R function, using the Euclidean distance. Clustering was then performed using the *hclust* R function using the complete method.

Analysis of isoform level expression and DEI-based isoform switching

Isoform level expression was quantified using IsoQuant[51] (v3.9.0) using default parameters, and low expression isoforms (isoforms where expression is ≥ 1 TPM in at least 3 samples) were further filtered to obtain a total of 279,998 isoforms (**Supplementary Table 2**). CPM values were converted to TPM values manually. EdgeR R-package[135] (v3.42.4) was also used on the isoform counts values to perform statistical analysis, as described in the **LrRNAseq data analysis of gene expression** section above. Differentially

expressed isoforms (DEIs) were established similar to DEGs, fold change ≥ 2 or ≤ 0.5, and FDR ≤ 0.05 when comparing reprogramming phases, but this time TPM values were used as expression values. This represented 198,247 isoforms that were selected for further analysis (**Supplementary Fig. 2** and **Supplementary Table 5**). For downstream analyses, only isoforms that were differentially expressed between D0 and the middle phase of reprogramming were selected. Finally, changes in isoform usage that were associated to the comparison between NPCs (D0) and D2 were kept only if they were not associated with the comparison between NPCs and wild-type NPCs placed in reprogramming medium for 2 days (FBSL) (i.e. with no reprogramming transgenes induction). This was done in order to select only differential isoform usage attributed to the effect of reprogramming transgenes expression and not due to media change. Isoforms were only selected if they were expressed in 40% of the samples. We obtained 35,694 isoforms that were subsequently used in Microsoft Excel to generate plots compiling the average expression in TPM of four categories (up-, down-, up- then down-, or down- then up-regulated) of DEIs between D0 and the middle phase of reprogramming (**Supplementary Table 5**). The differential usage of isoforms from a gene between D0 and the middle phase of reprogramming was used to identify genes that undergo isoform switch (IS) during this period according to an in-house matrix defined with Microsoft Excel (**Supplementary Table 5**). Observed percentages of genes undergoing IS among each biotype were calculated based on the total count of genes undergoing IS between D0 and the middle phase of reprogramming. Expected percentages were based on the total count of genes with DEIs during reprogramming, and p-values were calculated by Chi-squared test. Gene Ontology enrichment analysis of genes associated with at least one significant IS event per comparison was performed using DAVID[120] (v2022q3). Annotations were analyzed with a cut-off of FDR ≤ 0.1 for significant enrichment (**Supplementary Table 5**). Validation of *Snhg26* isoform switching between the NPC and middle phase samples was performed by PCR-amplifying from previous PCR-amplified libraries for 34 PCR cycles with ONT's forward cPRM and *Snhg26* isoforms' specific reverse primers (primer sequences are listed in **Supplementary Table 7**). Resulting amplicons were prepared for lrRNAseq and processed as in the **Library preparation and lrRNAseq, processing and alignment** section. For each sample, reads were

counted manually and averaged before dividing by the total number of mapped reads in the sample in order to get the expression level in CPM for both *Snhg26* isoforms (fastq reads archive is available in **Supplementary Data 5**).

Analysis of repetitive elements within splice junctions

First, a GTF file of every repetitive genomic sequence was created: 30nt sliding windows with 10nt steps were generated across the mouse mm10/GRCm38 genome with bedtools makewindows. Then, bedtools nuc was used to identify intervals where nucleotide composition was at least 85% GC, 85% AT or 85% of a single base. Overlapping repetitive windows were merged. Repetitive regions were merged to the annotated mm10 RepeatMasker .out file (available from UCSC: https://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/). The resulting repetitive regions that overlapped exons were extracted with bedtools intersect using our filtered transcriptome GTF file.

The effect of repetitive sequences on short-read RNA sequencing coverage was assessed on male and female mouse ESC datasets[136,137] (GEO accession: GSE67259) using Deeptools' bamCoverage with BPM normalization, ComputeMatrix using the exonic repetitive sequences bed file as regions file and the -b 250 -a 250 --binSize 1 --regionBodyLength 50 arguments and plotProfile using the --kmeans 1 --plotType se arguments (**Supplementary Fig. 3g**). A repetitive regions bed file is available as **Supplementary Data 2**.

Preparation of a catalog of validated isoform 5' ends

A bed file was generated from the refTSS version 4 transcript 5' ends dataset[138], which includes the FANTOM5 annotations, the TE-TSS database[139], 5' RNA sequencing of reprogramming cells from O'Malley *et al.*[129] and CAGEseq data from iPSCs[140]. To account for lower sequencing depths of the short-read sequencing within repetitive regions, we also included the repetitive regions obtained in the **Analysis of repetitive elements within splice junctions** section. The centre of each region was added as columns

7 and 8 to allow for calculation of distance to TSS by SQANTI3. The 5' ends bed file is available as **Supplementary Data 1**.

Analysis to support and validate isoform structure and splice junctions

Following isoform filtering (see **Mouse transcriptome assembly and isoform classification** section), isoforms were re-analyzed with SQANTI3 (v5.2.2) using the sqanti3_qc.py script, but this time included short-read and TSS support. Short-read datasets presented in **Supplementary Table 3** were downloaded as fastq files and supplied for isoform validation with the --short_reads argument to assess splice junctions support, short-read expression and TSS-ratio. The --CAGE_peak argument was used with the bed file generated in the **Preparation of a catalog of validated isoform 5' ends** section. The --polyA_motif_list argument was used with the default mouse poly(A) motifs provided in the SQANTI3 software.

To assess the presence of repetitive sequences within non-supported splice junctions, a bed file was generated from the SQANTI3 junctions.txt file corresponding to the 5' and 3' splice sites of each junction without short-reads support, surrounded by 150nt upstream of the 5' and downstream of the 3' splice site (150nt being the read length of most short-read datasets used for validation) (**Supplementary Fig. 3f**). The junctions bed file was intersected with the repetitive sequences file described in the **Analysis of repetitive elements within splice junctions** section. Junctions where at least one splice site overlapped a repetitive sequence were considered supported.

For coding isoforms with junctions remaining non-supported, the encoded protein sequences were compared to experimentally detected peptides obtained by mass spectrometry (see **Sample preparation for total proteome quantification**, **Mass spectrometry analysis**, and **DIA data analysis** sections below). First, isoforms with the same protein sequence as supported isoforms were excluded. The remaining isoforms for which peptides were detected were considered as supported.

Isoforms without TSS support by CAGEseq or comparable 5' sequencing datasets were assessed for their TSS-ratio (i.e. ratio of reads after/before the TSS). Based on thresholds established in Pardo-Palacios *et*

al.[54], isoforms with a TSS-ratio superior to 1.5 were considered to be supported. TSS were also compared to H3K4me3 peaks established in the **Analysis of publicly available ChIPseq data** section. A TSS lying within a H3K4me3 peak or within 100bp from the centre of the peak was considered supported. A classification file detailing support for all filtered isoforms is available as **Supplementary Table 2**.

Functional and structural annotation of isoforms

GFF3 files generated from SQANTI3's IsoAnnotLite pipeline were converted to a SQANTI3-like classification file (**Supplementary Data 4**). Untranslated region lengths of isoforms from genes subject to alternative splicing between reprogramming stages were compared with a non-parametric Kruskal Wallis test followed by Dunn's *post hoc* test. Percentage of non-coding or NMD isoforms from coding genes, or isoforms with PFAM domains were calculated per sample. Percentage of isoforms containing each annotation type was calculated per sample, and significant differences across reprogramming stages were assessed with an ANOVA followed by Tukey's *post hoc* test. Annotations are available in **Supplementary Table 2**.

Comparison to other transcriptome-building tools

Transcriptome was built using three orthogonal methods: Bambu[50] (v3.4.1) was used in R (v4.3.3) with a Novel Discovery Rate of 0.5 and the stranded = TRUE argument. IsoQuant[51] (v3.9.0) was used with the following arguments: --complete_genedb --sqanti_output --count_exons --check_canonical --report_canonical all --splice_correction_strategy conservative_ont --matching_strategy default --model_construction_strategy all --data_type nanopore --report_novel_unspliced true --polya_requirement never . Finally, StringTie[52] (v3.0.0) was run two ways: 1) with reference (Mus_musculus.GRCm38.99.gtf from Ensembl) and 2) *de novo*, both using the -s 1 -f 0 -L -v -G arguments. The resulting GTF files were then merged using StringTie with the --merge -i -f 0 -T 0 -F 0 -g 0 arguments.

Resulting GTF files were filtered to remove isoforms without strand information. Then, corresponding isoforms were identified using the sqanti3_qc.py script. The four generated GTF files were used

sequentially as references: (i) the GTF generated using the method described in the **Mouse transcriptome assembly and isoform classification** section, (ii) the Bambu GTF, (iii) the IsoQuant GTF, and (iv) the StringTie GTF. Each GTF file was compared against the others in a pairwise manner. Isoforms classified as Full Splice Matches were considered corresponding isoforms between two orthogonal methods. A list of corresponding isoforms is available as **Supplementary Table 3**. GTF files from orthogonal methods are available as **Supplementary Data 3**.

Analysis of alternative splicing

The occurrence of local alternative splicing (AS) and transcript events during reprogramming was assessed using SUPPA2[141] (v2.43) and the "suppa.py generateEvents" script with parameters "-f ioe -e SE SS MX RI FL" for AS events and parameter "-f ioi" for transcript events. A GTF file containing the expression filtered isoforms (i.e. 279,998 isoforms) was used as input for these analyses. To calculate the percent spliced in (PSI) values for each isoform and seven types of alternative splicing events, "suppa.py psiPerEvent" script was used with their respective events files (i.e. ".ioi" and ".ioe") and the TPM reads values from IsoQuant quantification above. Only the AS variations that had an adjusted p-value ≤ 0.05 in at least one reprogramming phase comparison were selected, reducing the list of 381,230 AS events to 164,341 AS events, and isoforms to 102,550 (events and isoforms are listed in **Supplementary Table 4**). Differential splicing between reprogramming phases was then established as isoforms with a $\Delta PSI \geq 0.1$ (more inclusion of an AS event) or $\Delta PSI \leq -0.1$ (less inclusion of an AS event) and an adjusted p-value ≤ 0.05. Total number of more or less included AS events for each type of AS was represented using Microsoft Excel. Differentially spliced isoforms between reprogramming phases were established as isoforms with a $\Delta PSI \geq 0.1$ (more abundant isoform) or $\Delta PSI \leq -0.1$ (less abundant isoform) and an adjusted p-value ≤ 0.05. Total number of alternatively spliced transcripts with more or less relative abundance was represented using Microsoft Excel. Similar to isoform usage calculations, AS and transcript events were kept only if they were associated to changes due mainly to reprogramming,

meaning event changes occurring due to media change as identified by comparing NPCs to FBSL were removed.

Overlap between genes with significant AS events (although not necessarily the same AS events) detected in this dataset (i.e. 10,490 genes) and other reprogramming datasets[32,35] was presented as a Venn diagram using InteractiVenn[142] (gene lists are presented in **Supplementary Table 4**). GO-term enrichment on genes commonly spliced was assessed using DAVID[120] (v2022q3) and FDR ≤ 0.1 (**Supplementary Table 4**).

FYN target genes were selected using the CRISPR screen, RNAi screen and Predictability datasets from the DepMap Public 25Q3 Release dataset through the DepMap portal[68] (https://depmap.org/portal). Their expression was averaged depending on whether they correlated negatively or positively with *Fyn* expression in the CRISPR screen, based on a correlation threshold ≥ 0.3 or ≤ -0.3 (gene lists are listed in **Supplementary Table 4**).

Visualization of gene loci and alternatively spliced isoforms

Validations of gene loci, alternative splicing events, isoform abundance and isoform switching were performed using the Integrative Genomics Viewer (IGV)[143] (v2.16.2) tracks.

Identification of candidate lncRNAs

A first list of genes for which alternatively spliced isoforms were significantly more or less relatively abundant (i.e. IS analysis based on ΔPSI events method, see **Analysis of alternative splicing** section) between D0 and the middle phase of reprogramming, removing those occurring when comparing NPCs to FBSL samples, which represent changes due to media, was created (i.e. 12,086 genes). A second list of genes that undergo isoform switching (IS analysis based on DEIs method, see **Analysis of isoform level expression and isoform switching** section), between D0 and the middle phase of reprogramming, and which are not due to media change, was generated (i.e. 2,374 genes). The two lists are presented in **Supplementary Table 6**. Overlap between these two lists (i.e. 1,766 genes) was presented as a Venn

diagram using Microsoft Excel. LncRNAs from this overlap were cross-referenced with lncRNAs detected in Hussein *et al.*[16,41].

Expression of candidate lncRNAs from publicly available datasets (all listed in **Supplementary Table 7**) were represented with SEM using Microsoft Excel or GraphPad Prism (v7.04). P-values between the frequencies were calculated using t-test (unpaired and unequal variance).

Sample preparation for total proteome quantification

25µg of cell samples were processed per sample type. Each sample was reduced with DTT at 10mM for 30 minutes at 56°C and alkylated with iodoacetamide at 30mM for 45 minutes at room temperature in the dark. Each sample was then brought up to a final volume of 50L. 6.25µL of MagReSyn HILIC beads (Resin Biosciences; Gauteng, South Africa) per sample were transferred to a microcentrifuge tube and washed 2 times with 200µL Equilibration buffer (15% ACN in 100mM Ammonium acetate, pH 4.5). Each 12.5µL initial bead slurry was moved into a new microcentrifuge tube in equilibration buffer. 50µL Binding buffer (30% ACN in 200mM Ammonium acetate, pH 4.5) was added to 50µL protein solution for a final volume of 100µL. Equilibration buffer was removed from the beads and 100µL protein solution was added and mixed at 1,200 rpm for 30 minutes at room temperature. Beads were washed twice with 200µL Wash buffer (95% ACN), mixed for 1 minute at 1,200rpm. All wash buffer was removed and beads resuspended in 100µL digestion buffer (50mM Tris-HCl, pH 8.0) with 1µg trypsin/Lys-C. Protein was digested for 1 hour at 46°C with mixing (1,200 rpm). Peptides were collected and acidified to 2% formic acid. Peptides were dried down and stored at -40°C until mass spectrometry (MS) acquisition.

Mass spectrometry analysis

For data-independent acquisition (DIA) liquid chromatography-Tandem mass spectrometry (LC-MS/MS), 500ng protein equivalent of digested peptides were analyzed using a nano-high-performance liquid chromatography (HPLC) coupled to MS. Samples were separated on an Aurora Elite TS column (IonOpticks Pty Ltd. #AUR3-15075C18-TS). The sample in 5% formic acid was trap (300µm i.d., 0.5cm

length, #174500) loaded at 800 bar, onto a 75µm i.d. x 15cm nano-spray emitter (packed with 1.7µm C18 beads) heated at 50°C with the TS Interface. Peptides were eluted from the column with an acetonitrile gradient generated by an Vanquish Neo UHPLC System (ThermoFisher Scientific), and analyzed on an Orbitrap™ Astral™. The gradient was delivered at 800nL/min from 0.8% acetonitrile with 0.1% formic acid to 3.2% acetonitrile with 0.1% formic acid over 1 minute, 3.2% to 6.4% acetonitrile with 0.1% formic acid over 1 minute, and 6.4% to 35.2% acetonitrile with 0.1% formic acid over 19 minutes. This was followed by a column wash of 76% acetonitrile with 0.1% formic acid over 2 minutes. The LC method ends with an equilibration of 2 column volumes at the same combined control as loading. The total DIA protocol is 23 minutes. Advanced peak determination was turned on with expected LC peak width of 12 seconds. FAIMS was used with -50V compensation voltage. MS1 scan was from 380-980m/z with an orbitrap resolution of 240,000, normalized AGC target of 500%, and 50ms maximum injection time in profile mode. DIA windows were set to auto with 4m/z window sizes with 0 overlap, in the precursor range of 380-980m/z. HCD collision energy was set to 25% using the Astral detector. The scan range was 150-2000m/z with RF lens set to 50%, AGC target to 500%, and maximum injection time of 3 ms.

DIA data analysis

Spectronaut v20 directDIA+ (Biognosis AG; Switzerland) BGS workflow was used to search the data with the Spectronaut generated mouse spectral library (Mouse_Uniprot_2023_06_13) supplemented with the fasta file derived from the SQANTI3 output files, as defined in the **Mouse transcriptome assembly and isoform classification** section. Parameters for the search were default, normalization was on with runwise imputation. Differential abundance testing used was unpaired t-test (data is available in **Supplementary Table 3**).

Inclusion & Ethics statement

All collaborators of this study that fulfilled Nature Portfolio journals' criteria for authorship have been included as authors, as their participation was essential for the design and implementation of the study. Roles and responsibilities were agreed among collaborators ahead of the research. This research would not have been severely restricted or prohibited in the setting of the researchers.

Reporting summary

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

Data availability

Datasets from other referenced works can be accessed either in the NCBI Gene Expression Omnibus (GEO) repository using their GSE numbers or in the database of the EMBL's European Bioinformatics Institute using their E-MTAB numbers (all reference datasets and processed datasets with links and descriptions are summarized in **Supplementary Tables 3**, **6** and **7**). Long-read RNA sequencing raw data, filtered aligned reads and processed data, including mouse transcriptome annotation and expression level matrices, from this study are publicly available and have been deposited in the GEO database. All mass spectrometry files used in this study were deposited in MassIVE (http://massive.ucsd.edu). All cell lines and biological materials generated in this study are available from the corresponding author upon direct request.

Acknowledgements

We thank Dr. Ian Roger and Dr. Sophie Mockly for advice and reading of the manuscript. Gaëlle Bourriquen assisted with the mouse CRISPRi sgRNA design. We also acknowledge the contribution of the Model Production and Pathology Cores at The Centre for Phenogenomics for technical support in generating chimeras and processing teratomas. We thank the Network Biology Collaborative Centre Proteomics Facility (RRID: SCR_025375) at the Lunenfeld-Tanenbaum Research Institute for total proteome analysis. The facility is supported by the Canada Foundation for Innovation and the Ontario Government. This work was funded by the Canadian Institutes of Health Research (CIHR) to S.M.I.H. (PJT-378019) and supported by the Fonds de recherche du Québec (FRQ) through the Research Centre Grant for the CHU de Québec-Université Laval Research Centre (30641). V.F. is a recipient of a training award of the Fonds de recherche du Québec - Santé (FRQS) https://doi.org/10.69777/291836. G.K. is a recipient of training awards form the Natural Sciences and Engineering Research Council of Canada (NSERC) and the FRQS https://doi.org/10.69777/307170. V.W. is a recipient of a training award of the Fonds de recherche du Québec - Nature et Technologies (FRQNT) https://doi.org/10.69777/271653. S.M.I.H. and J.P.L. are Junior 2 Research Scholars of the FRQS https://doi.org/10.69777/310750 and https://doi.org/10.69777/312369.

Author Contributions

V.F., J.P.L., Y.Y., J.W., and S.M.I.H. conceived of and designed the study. V.F., G.K., R.H., V.W., M.P., N.Y., M.G., E.I.J.L., J.L., and J.P.L. developed and performed the experiments or collected data. V.F., G.K., J.P.L, and S.M.I.H. designed and performed the computational and statistical analyses. V.F., G.K., R.H., V.W., M.G., E.I.J.L., V.M., J.P.L., Y.Y, J.W., and S.M.I.H. wrote, reviewed and edited the manuscript.

Competing Interests

The authors declare no competing interests.

Figure 1

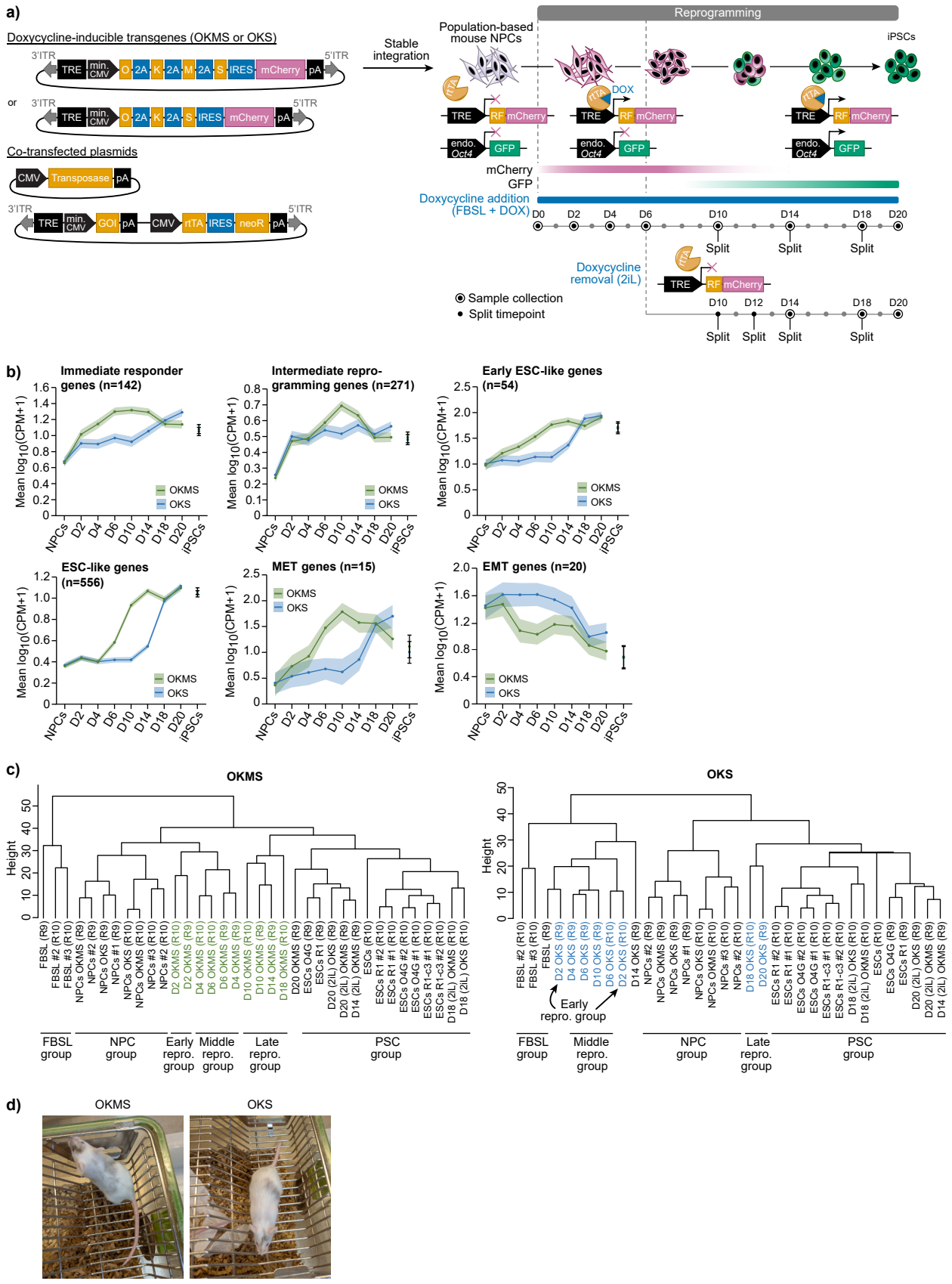

a)
Doxycycline-inducible transgenes (OKMS or OKS)
3'ITR
TRE
min. CMV
O
2A
K
M
S
IRES
mCherry
pA
5'ITR
or
Co-transfected plasmids
CMV
Transposase
GOI
rtTA
neoR
Stable integration
Reprogramming
Population-based mouse NPCs
iPSCs
DOX
RF
endo. Oct4
GFP
mCherry
Doxycycline addition (FBSL + DOX)
D0
D2
D4
D6
D10
D14
D18
D20
Split
Doxycycline removal (2iL)
D12
Sample collection
Split timepoint
b)
Immediate responder genes (n=142)
Intermediate reprogramming genes (n=271)
Early ESC-like genes (n=54)
ESC-like genes (n=556)
MET genes (n=15)
EMT genes (n=20)
Mean log10(CPM+1)
OKMS
OKS
NPCs
iPSCs
c)
OKMS
OKS
Height
Early repro. group
FBSL group
NPC group
Middle repro. group
Late repro. group
PSC group
d)
OKMS
OKS

Figure 2

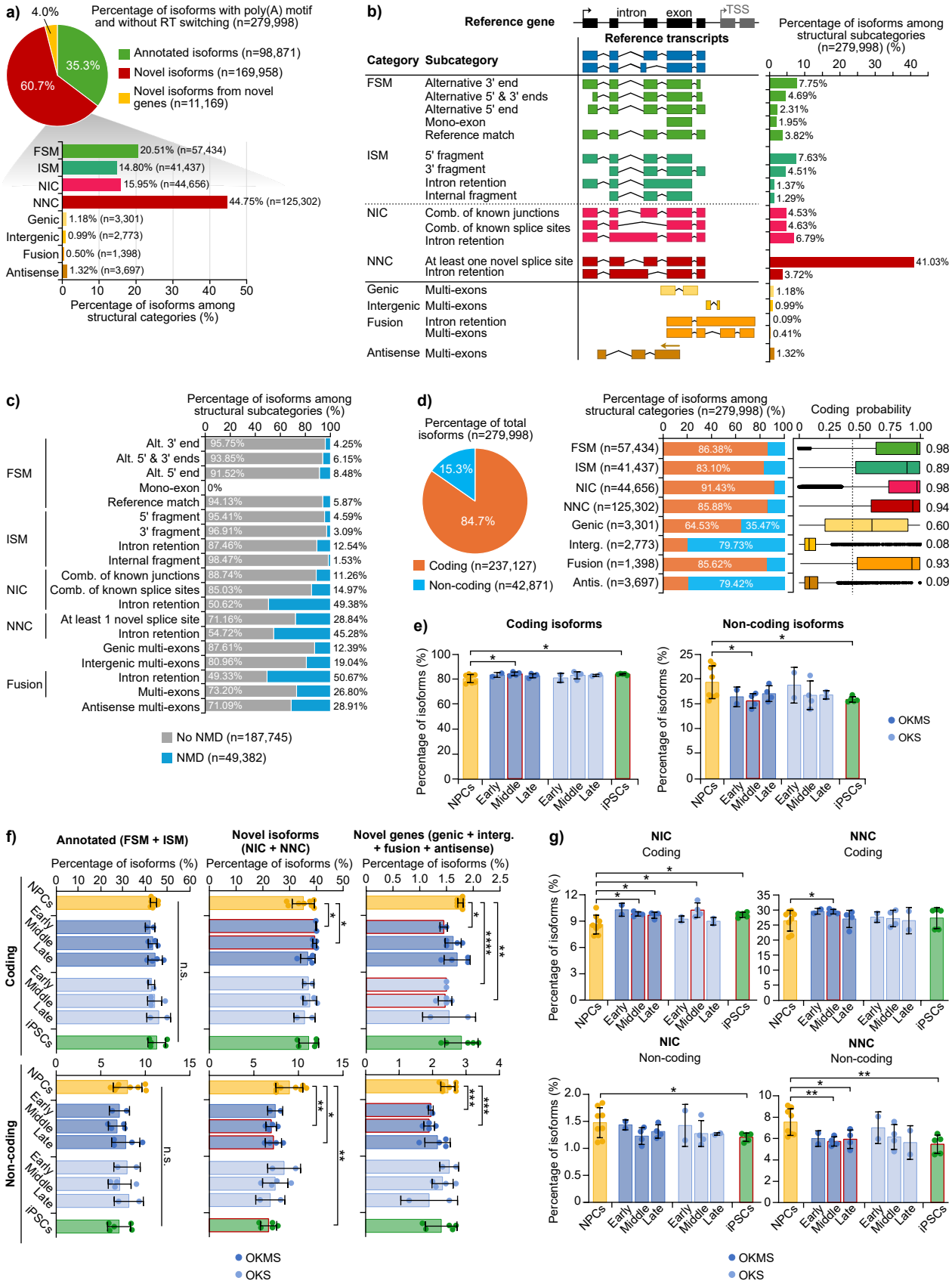

a)
Percentage of isoforms with poly(A) motif and without RT switching (n=279,998)
4.0%
35.3%
60.7%
Annotated isoforms (n=98,871)
Novel isoforms (n=169,958)
Novel isoforms from novel genes (n=11,169)
FSM 20.51% (n=57,434)
ISM 14.80% (n=41,437)
NIC 15.95% (n=44,656)
NNC 44.75% (n=125,302)
Genic 1.18% (n=3,301)
Intergenic 0.99% (n=2,773)
Fusion 0.50% (n=1,398)
Antisense 1.32% (n=3,697)
Percentage of isoforms among structural categories (%)
b)
Reference gene
intron
exon
TSS
Reference transcripts
Category
Subcategory
Percentage of isoforms among structural subcategories (n=279,998) (%)
FSM
Alternative 3' end 7.75%
Alternative 5' & 3' ends 4.69%
Alternative 5' end 2.31%
Mono-exon 1.95%
Reference match 3.82%
ISM
5' fragment 7.63%
3' fragment 4.51%
Intron retention 1.37%
Internal fragment 1.29%
NIC
Comb. of known junctions 4.53%
Comb. of known splice sites 4.63%
Intron retention 6.79%
NNC
At least one novel splice site 41.03%
Intron retention 3.72%
Genic Multi-exons 1.18%
Intergenic Multi-exons 0.99%
Fusion Intron retention 0.09%
Multi-exons 0.41%
Antisense Multi-exons 1.32%
c)
Percentage of isoforms among structural subcategories (%)
FSM
Alt. 3' end 95.75% 4.25%
Alt. 5' & 3' ends 93.85% 6.15%
Alt. 5' end 91.52% 8.48%
Mono-exon 0%
Reference match 94.13% 5.87%
ISM
5' fragment 95.41% 4.59%
3' fragment 96.91% 3.09%
Intron retention 87.46% 12.54%
Internal fragment 98.47% 1.53%
NIC
Comb. of known junctions 88.74% 11.26%
Comb. of known splice sites 85.03% 14.97%
Intron retention 50.62% 49.38%
NNC
At least 1 novel splice site 71.16% 28.84%
Intron retention 54.72% 45.28%
Genic multi-exons 87.61% 12.39%
Intergenic multi-exons 80.96% 19.04%
Fusion
Intron retention 49.33% 50.67%
Multi-exons 73.20% 26.80%
Antisense multi-exons 71.09% 28.91%
No NMD (n=187,745)
NMD (n=49,382)
d)
Percentage of total isoforms (n=279,998)
15.3%
84.7%
Coding (n=237,127)
Non-coding (n=42,871)
Percentage of isoforms among structural categories (n=279,998) (%)
Coding probability
FSM (n=57,434) 86.38% 0.98
ISM (n=41,437) 83.10% 0.89
NIC (n=44,656) 91.43% 0.98
NNC (n=125,302) 85.88% 0.94
Genic (n=3,301) 64.53% 35.47% 0.60
Interg. (n=2,773) 79.73% 0.08
Fusion (n=1,398) 85.62% 0.93
Antis. (n=3,697) 79.42% 0.09
e)
Coding isoforms
Non-coding isoforms
Percentage of isoforms (%)
NPCs
Early
Middle
Late
iPSCs
OKMS
OKS
f)
Annotated (FSM + ISM)
Novel isoforms (NIC + NNC)
Novel genes (genic + interg. + fusion + antisense)
Percentage of isoforms (%)
Coding
Non-coding
n.s.
OKMS
OKS
g)
NIC
Coding
NNC
Coding
NIC
Non-coding
NNC
Non-coding
Percentage of isoforms (%)
NPCs
Early
Middle
Late
iPSCs
OKMS
OKS

Figure 3

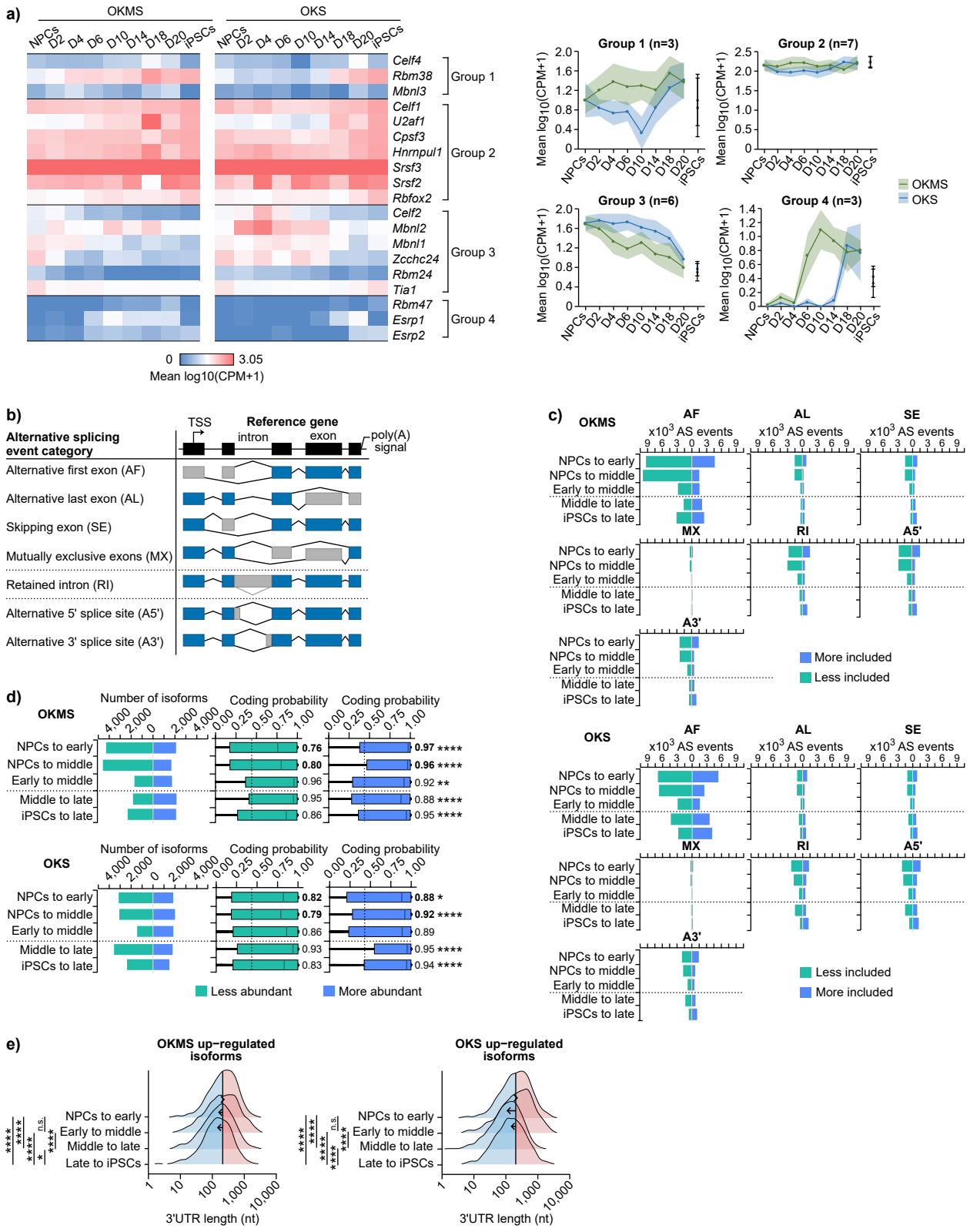

a)
OKMS
OKS
NPCs D2 D4 D6 D10 D14 D18 D20 iPSCs
Celf4
Rbm38
Mbnl3
Group 1
Celf1
U2af1
Cpsf3
Hnrnpul1
Srsf3
Srsf2
Rbfox2
Group 2
Celf2
Mbnl2
Mbnl1
Zcchc24
Rbm24
Tia1
Group 3
Rbm47
Esrp1
Esrp2
Group 4
0
3.05
Mean log10(CPM+1)
Group 1 (n=3)
Group 2 (n=7)
Group 3 (n=6)
Group 4 (n=3)
Mean log10(CPM+1)
OKMS
OKS
b)
Alternative splicing event category
TSS
Reference gene
intron
exon
poly(A) signal
Alternative first exon (AF)
Alternative last exon (AL)
Skipping exon (SE)
Mutually exclusive exons (MX)
Retained intron (RI)
Alternative 5' splice site (A5')
Alternative 3' splice site (A3')
c)
OKMS
AF
AL
SE
x10^3 AS events
NPCs to early
NPCs to middle
Early to middle
Middle to late
iPSCs to late
MX
RI
A5'
A3'
More included
Less included
OKS
d)
OKMS
Number of isoforms
Coding probability
Coding probability
NPCs to early
NPCs to middle
Early to middle
Middle to late
iPSCs to late
0.76
0.80
0.96
0.95
0.86
0.97 ****
0.96 ****
0.92 **
0.88 ****
0.95 ****
OKS
0.82
0.79
0.86
0.93
0.83
0.88 *
0.92 ****
0.89
0.95 ****
0.94 ****
Less abundant
More abundant
e)
OKMS up-regulated isoforms
OKS up-regulated isoforms
NPCs to early
Early to middle
Middle to late
Late to iPSCs
3'UTR length (nt)

Figure 4

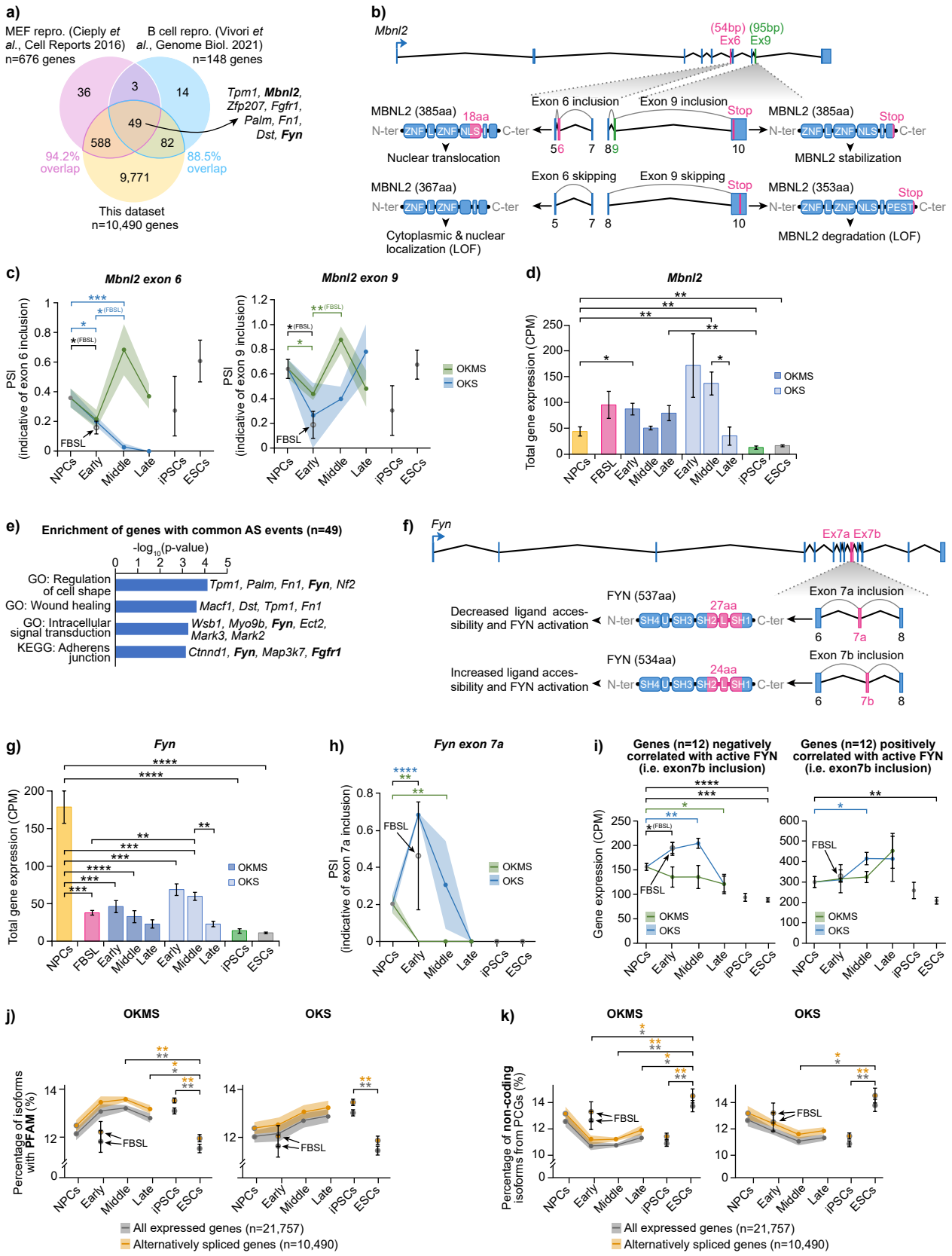
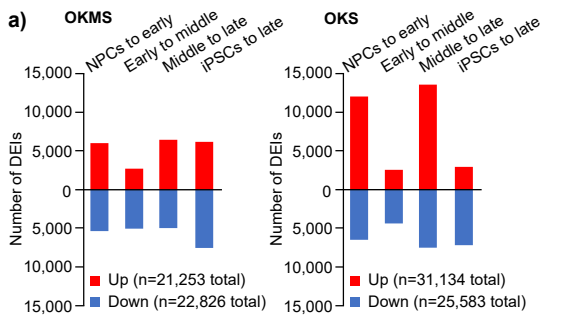
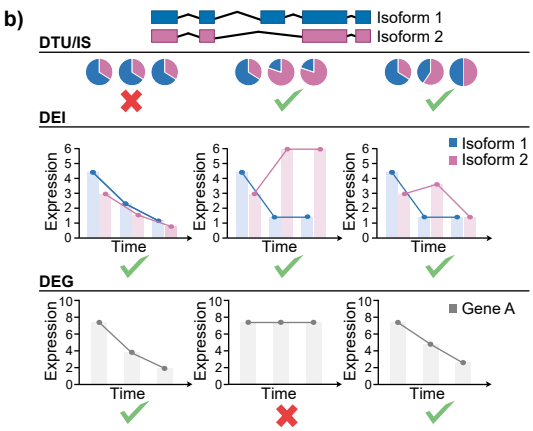
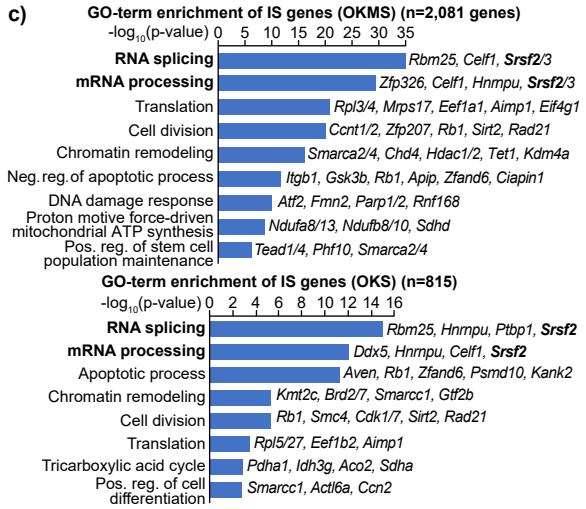
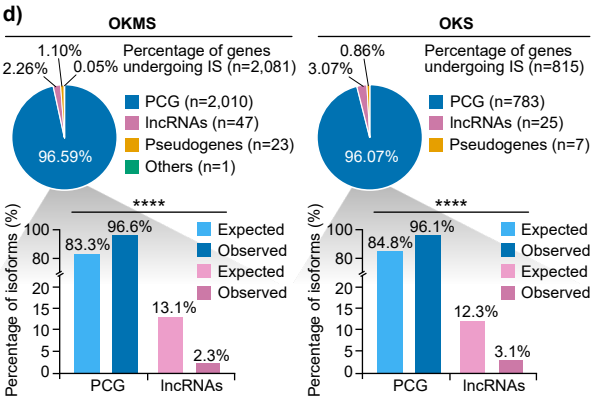

a)
MEF repro. (Cieply et al., Cell Reports 2016)
n=676 genes
B cell repro. (Vivori et al., Genome Biol. 2021)
n=148 genes
36
3
14
49
588
82
9,771
94.2% overlap
88.5% overlap
Tpm1, Mbnl2, Zfp207, Fgfr1, Palm, Fn1, Dst, Fyn
This dataset
n=10,490 genes
b)
Mbnl2
(54bp) Ex6
(95bp) Ex9
Exon 6 inclusion
Exon 9 inclusion
MBNL2 (385aa)
18aa
N-ter
C-ter
Nuclear translocation
Stop
MBNL2 stabilization
Exon 6 skipping
Exon 9 skipping
MBNL2 (367aa)
Cytoplasmic & nuclear localization (LOF)
MBNL2 (353aa)
PEST
MBNL2 degradation (LOF)
c)
Mbnl2 exon 6
Mbnl2 exon 9
PSI (indicative of exon 6 inclusion)
PSI (indicative of exon 9 inclusion)
FBSL
OKMS
OKS
NPCs
Early
Middle
Late
iPSCs
ESCs
d)
Mbnl2
Total gene expression (CPM)
e)
Enrichment of genes with common AS events (n=49)
-log10(p-value)
GO: Regulation of cell shape
Tpm1, Palm, Fn1, Fyn, Nf2
GO: Wound healing
Macf1, Dst, Tpm1, Fn1
GO: Intracellular signal transduction
Wsb1, Myo9b, Fyn, Ect2, Mark3, Mark2
KEGG: Adherens junction
Ctnnd1, Fyn, Map3k7, Fgfr1
f)
Fyn
Ex7a Ex7b
Exon 7a inclusion
Exon 7b inclusion
FYN (537aa)
FYN (534aa)
27aa
24aa
Decreased ligand accessibility and FYN activation
Increased ligand accessibility and FYN activation
SH4
SH3
SH2
SH1
7a
7b
g)
Fyn
Total gene expression (CPM)
h)
Fyn exon 7a
PSI (indicative of exon 7a inclusion)
i)
Genes (n=12) negatively correlated with active FYN (i.e. exon7b inclusion)
Genes (n=12) positively correlated with active FYN (i.e. exon7b inclusion)
Gene expression (CPM)
j)
OKMS
OKS
Percentage of isoforms with PFAM (%)
k)
Percentage of non-coding isoforms from PCGs (%)
All expressed genes (n=21,757)
Alternatively spliced genes (n=10,490)

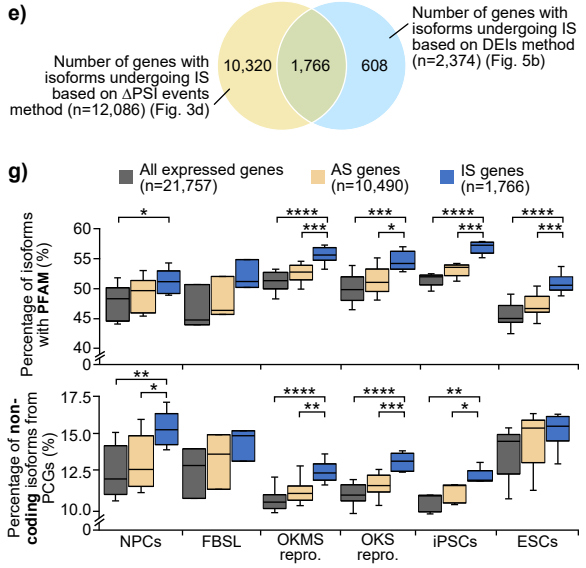
Figure 5

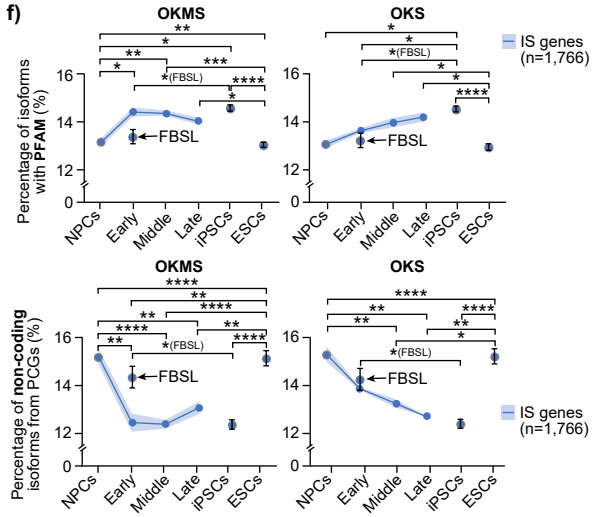

a)
OKMS
OKS
NPCs to early
Early to middle
Middle to late
iPSCs to late
Number of DEIs
Up (n=21,253 total)
Down (n=22,826 total)
Up (n=31,134 total)
Down (n=25,583 total)


b)
DTU/IS
Isoform 1
Isoform 2
DEI
Expression
Time
DEG
Gene A

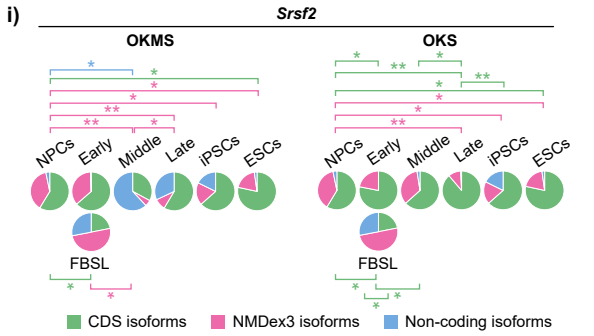

c)
GO-term enrichment of IS genes (OKMS) (n=2,081 genes)
-log10(p-value)
RNA splicing
mRNA processing
Translation
Cell division
Chromatin remodeling
Neg. reg. of apoptotic process
DNA damage response
Proton motive force-driven mitochondrial ATP synthesis
Pos. reg. of stem cell population maintenance
Rbm25, Celf1, Srsf2/3
Zfp326, Celf1, Hnrnpu, Srsf2/3
Rpl3/4, Mrps17, Eef1a1, Aimp1, Eif4g1
Ccnt1/2, Zfp207, Rb1, Sirt2, Rad21
Smarca2/4, Chd4, Hdac1/2, Tet1, Kdm4a
Itgb1, Gsk3b, Rb1, Apip, Zfand6, Ciapin1
Atf2, Fmn2, Parp1/2, Rnf168
Ndufa8/13, Ndufb8/10, Sdhd
Tead1/4, Phf10, Smarca2/4
GO-term enrichment of IS genes (OKS) (n=815)
-log10(p-value)
RNA splicing
mRNA processing
Apoptotic process
Chromatin remodeling
Cell division
Translation
Tricarboxylic acid cycle
Pos. reg. of cell differentiation
Rbm25, Hnrnpu, Ptbp1, Srsf2
Ddx5, Hnrnpu, Celf1, Srsf2
Aven, Rb1, Zfand6, Psmd10, Kank2
Kmt2c, Brd2/7, Smarcc1, Gtf2b
Rb1, Smc4, Cdk1/7, Sirt2, Rad21
Rpl5/27, Eef1b2, Aimp1
Pdha1, Idh3g, Aco2, Sdha
Smarcc1, Actl6a, Ccn2


d)
OKMS
OKS
Percentage of genes undergoing IS (n=2,081)
Percentage of genes undergoing IS (n=815)
1.10%
2.26%
0.05%
96.59%
0.86%
3.07%
96.07%
PCG (n=2,010)
lncRNAs (n=47)
Pseudogenes (n=23)
Others (n=1)
PCG (n=783)
lncRNAs (n=25)
Pseudogenes (n=7)
Percentage of isoforms (%)
Expected
Observed
83.3%
96.6%
13.1%
2.3%
84.8%
96.1%
12.3%
3.1%
PCG
lncRNAs


e)
Number of genes with isoforms undergoing IS based on ΔPSI events method (n=12,086) (Fig. 3d)
10,320
1,766
608
Number of genes with isoforms undergoing IS based on DEIs method (n=2,374) (Fig. 5b)


f)
OKMS
OKS
IS genes (n=1,766)
Percentage of isoforms with PFAM (%)
Percentage of non-coding isoforms from PCGs (%)
FBSL
NPCs
Early
Middle
Late
iPSCs
ESCs

g)
All expressed genes (n=21,757)
AS genes (n=10,490)
IS genes (n=1,766)
Percentage of isoforms with PFAM (%)
Percentage of non-coding isoforms from PCGs (%)
NPCs
FBSL
OKMS repro.
OKS repro.
iPSCs
ESCs


i)
Srsf2
OKMS
OKS
NPCs
Early
Middle
Late
iPSCs
ESCs
FBSL
CDS isoforms
NMDex3 isoforms
Non-coding isoforms

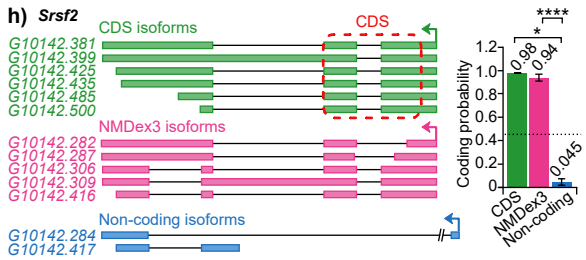

h)
Srsf2
CDS isoforms
CDS
G10142.381
G10142.399
G10142.425
G10142.435
G10142.485
G10142.500
NMDex3 isoforms
G10142.282
G10142.287
G10142.306
G10142.309
G10142.416
Non-coding isoforms
G10142.284
G10142.417
Coding probability
0.98
0.94
0.045
CDS
NMDex3
Non-coding

Figure 6

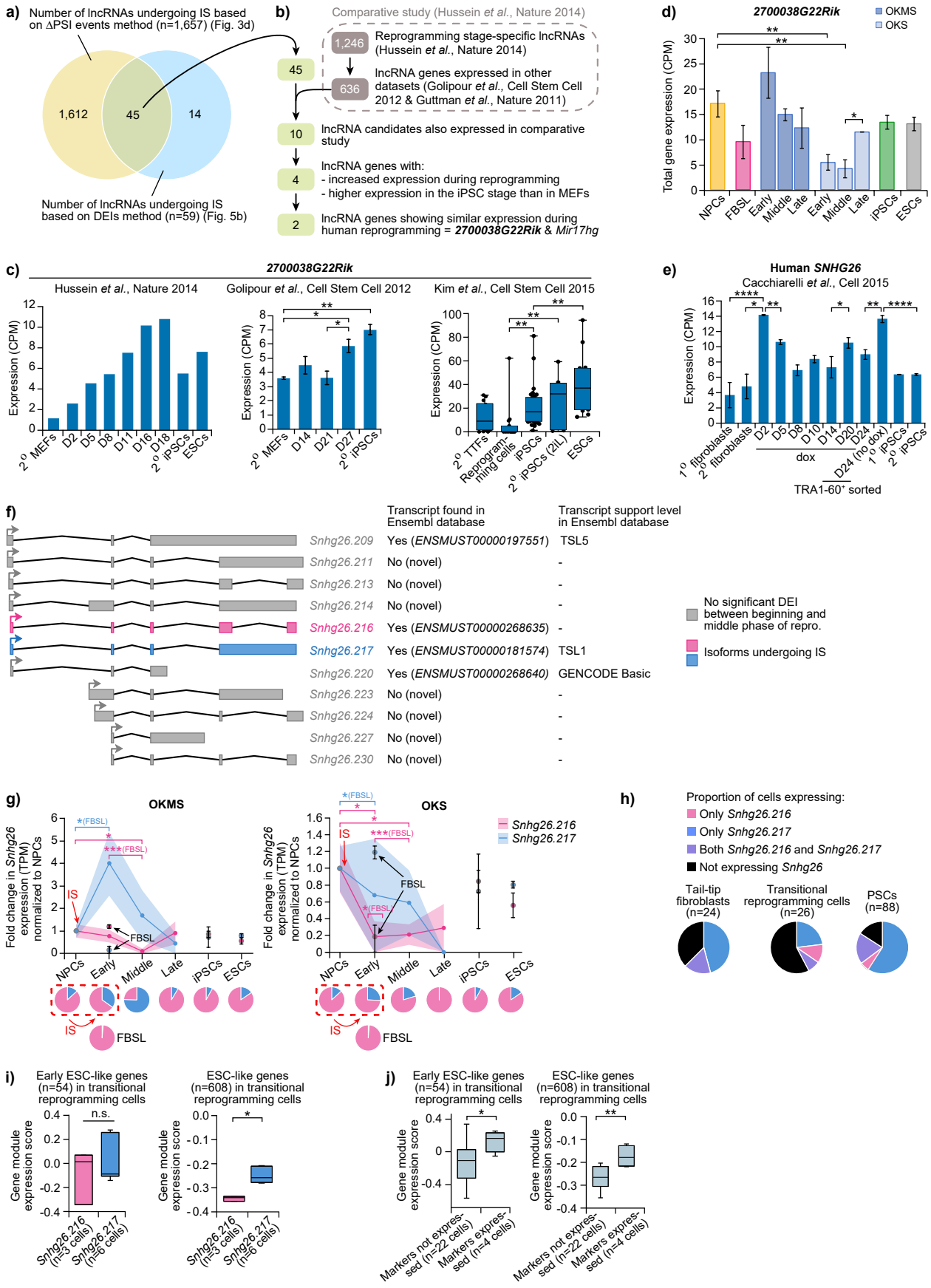

a)
Number of lncRNAs undergoing IS based on ΔPSI events method (n=1,657) (Fig. 3d)
1,612
45
14
Number of lncRNAs undergoing IS based on DEIs method (n=59) (Fig. 5b)
b)
Comparative study (Hussein et al., Nature 2014)
1,246
Reprogramming stage-specific lncRNAs (Hussein et al., Nature 2014)
636
lncRNA genes expressed in other datasets (Golipour et al., Cell Stem Cell 2012 & Guttman et al., Nature 2011)
45
10
lncRNA candidates also expressed in comparative study
4
lncRNA genes with:
- increased expression during reprogramming
- higher expression in the iPSC stage than in MEFs
2
lncRNA genes showing similar expression during human reprogramming = 2700038G22Rik & Mir17hg
d)
2700038G22Rik
OKMS
OKS
Total gene expression (CPM)
NPCs
FBSL
Early
Middle
Late
Early
Middle
Late
iPSCs
ESCs
c)
2700038G22Rik
Hussein et al., Nature 2014
Golipour et al., Cell Stem Cell 2012
Kim et al., Cell Stem Cell 2015
Expression (CPM)
2° MEFs
D2
D5
D8
D11
D16
D18
2° iPSCs
ESCs
D14
D21
D27
2° TTFs
Reprogram-ming cells
iPSCs
2° iPSCs (2iL)
e)
Human SNHG26
Cacchiarelli et al., Cell 2015
Expression (CPM)
1° fibroblasts
2° fibroblasts
D2
D5
D8
D10
D14
D20
D24
D24 (no dox)
1° iPSCs
2° iPSCs
dox
TRA1-60+ sorted
f)
Transcript found in Ensembl database
Transcript support level in Ensembl database
Snhg26.209 Yes (ENSMUST00000197551) TSL5
Snhg26.211 No (novel) -
Snhg26.213 No (novel) -
Snhg26.214 No (novel) -
Snhg26.216 Yes (ENSMUST00000268635) -
Snhg26.217 Yes (ENSMUST00000181574) TSL1
Snhg26.220 Yes (ENSMUST00000268640) GENCODE Basic
Snhg26.223 No (novel) -
Snhg26.224 No (novel) -
Snhg26.227 No (novel) -
Snhg26.230 No (novel) -
No significant DEI between beginning and middle phase of repro.
Isoforms undergoing IS
g)
OKMS
OKS
Fold change in Snhg26 expression (TPM) normalized to NPCs
*(FBSL)
***(FBSL)
IS
FBSL
Snhg26.216
Snhg26.217
NPCs
Early
Middle
Late
iPSCs
ESCs
h)
Proportion of cells expressing:
Only Snhg26.216
Only Snhg26.217
Both Snhg26.216 and Snhg26.217
Not expressing Snhg26
Tail-tip fibroblasts (n=24)
Transitional reprogramming cells (n=26)
PSCs (n=88)
i)
Early ESC-like genes (n=54) in transitional reprogramming cells
ESC-like genes (n=608) in transitional reprogramming cells
Gene module expression score
n.s.
Snhg26.216 (n=3 cells)
Snhg26.217 (n=6 cells)
j)
Early ESC-like genes (n=54) in transitional reprogramming cells
ESC-like genes (n=608) in transitional reprogramming cells
Gene module expression score
Markers not expressed (n=22 cells)
Markers expressed (n=4 cells)

Figure 7

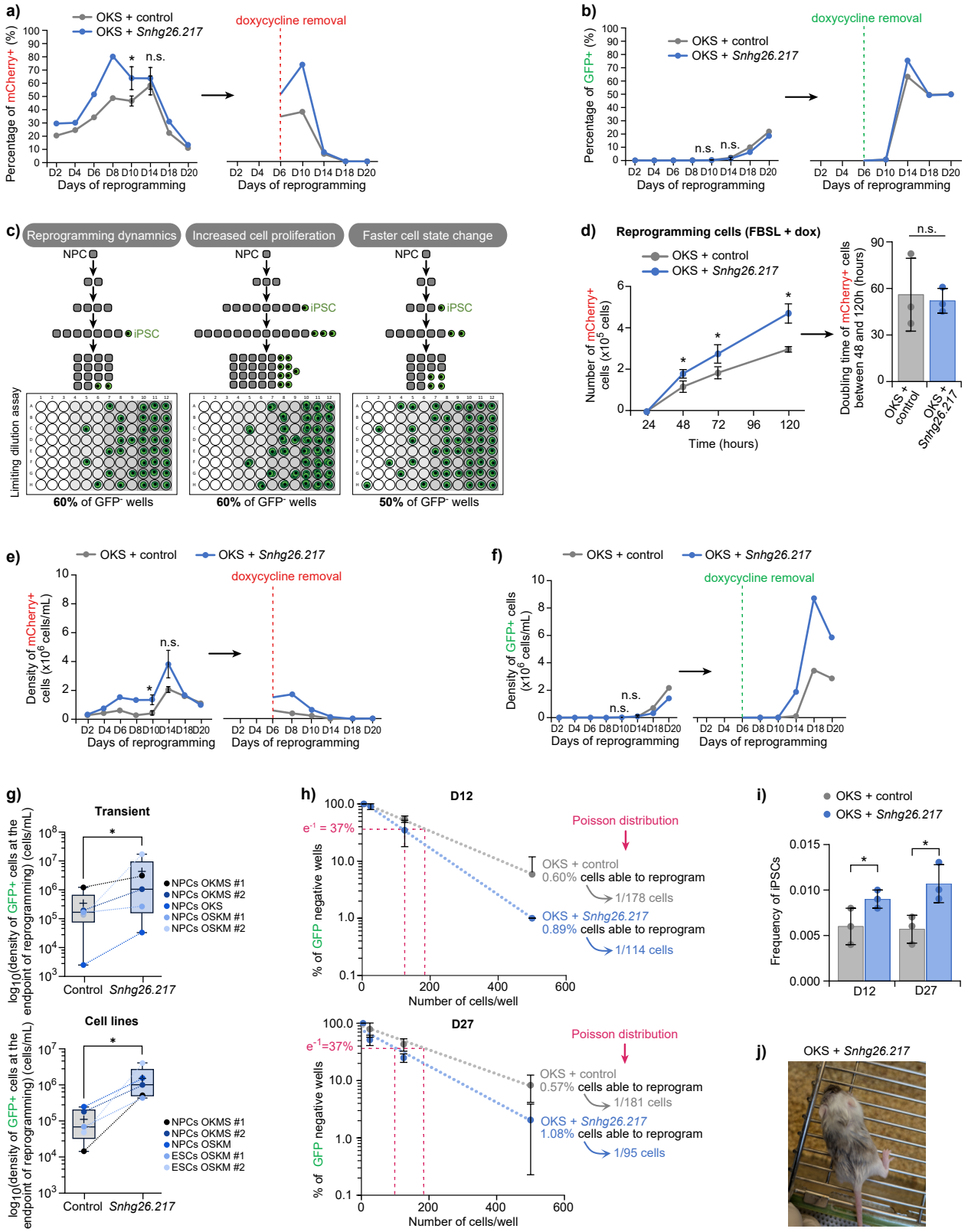

a)
OKS + control
OKS + Snhg26.217
doxycycline removal
Percentage of mCherry+ (%)
Days of reprogramming
b)
Percentage of GFP+ (%)
c)
Reprogramming dynamnics
Increased cell proliferation
Faster cell state change
NPC
iPSC
Limiting dilution assay
60% of GFP- wells
60% of GFP- wells
50% of GFP- wells
d)
Reprogramming cells (FBSL + dox)
Number of mCherry+ cells (x10^5 cells)
Time (hours)
Doubling time of mCherry+ cells between 48 and 120h (hours)
OKS + control
OKS + Snhg26.217
e)
Density of mCherry+ cells (x10^6 cells/mL)
f)
Density of GFP+ cells (x10^6 cells/mL)
g)
Transient
Cell lines
log10(density of GFP+ cells at the endpoint of reprogramming) (cells/mL)
NPCs OKMS #1
NPCs OKMS #2
NPCs OKS
NPCs OSKM #1
NPCs OSKM #2
NPCs OKSM
ESCs OSKM #1
ESCs OSKM #2
Control
Snhg26.217
h)
D12
D27
Poisson distribution
% of GFP negative wells
Number of cells/well
OKS + control
0.60% cells able to reprogram
1/178 cells
OKS + Snhg26.217
0.89% cells able to reprogram
1/114 cells
0.57% cells able to reprogram
1/181 cells
1.08% cells able to reprogram
1/95 cells
i)
Frequency of iPSCs
D12
D27
j)
OKS + Snhg26.217

Figure 8

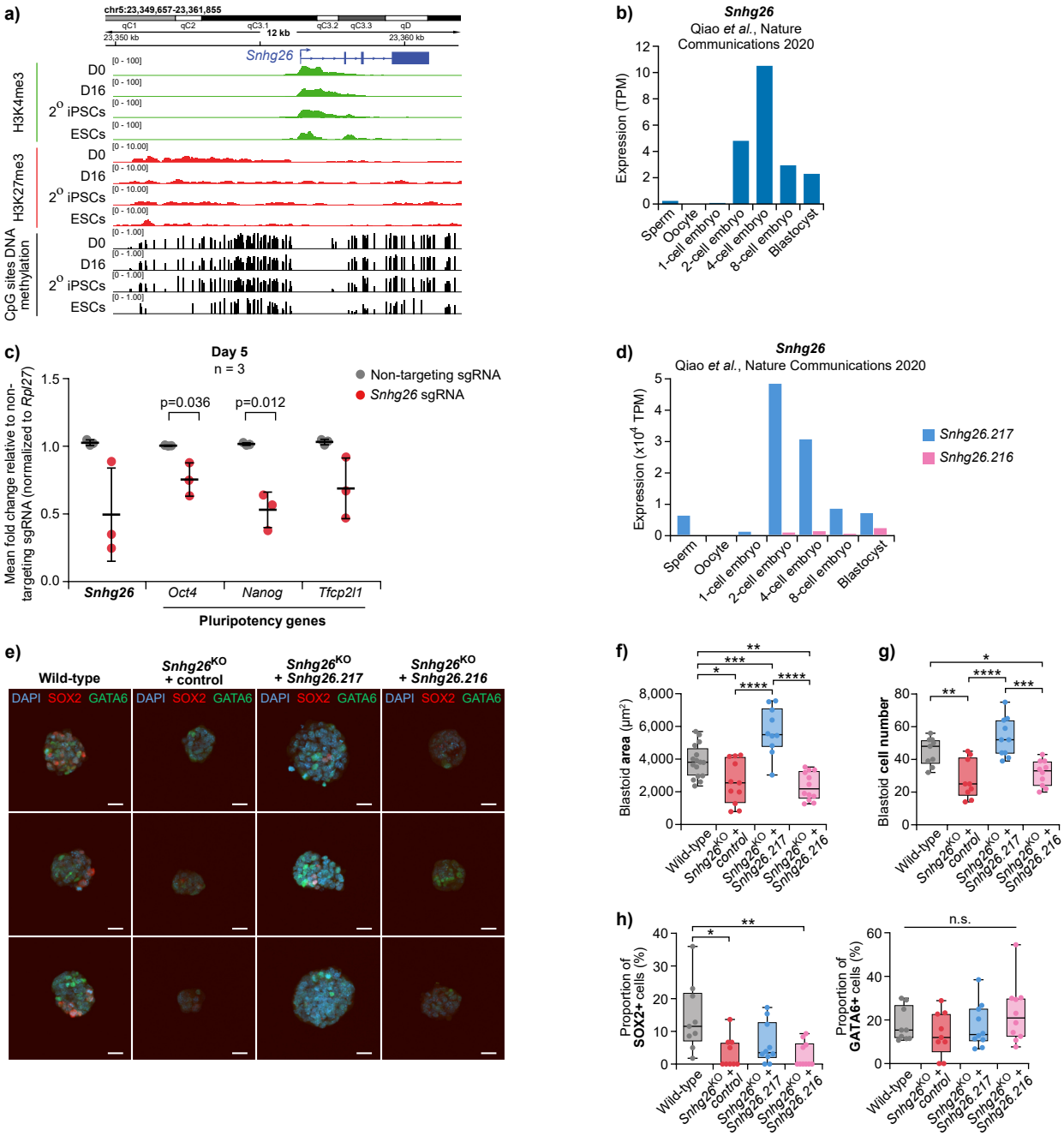

a)
chr5:23,349,657-23,361,855
qC1
qC2
qC3.1
qC3.2
qC3.3
qD
12 kb
23,350 kb
23,360 kb
Snhg26
H3K4me3
H3K27me3
CpG sites DNA methylation
D0
D16
2° iPSCs
ESCs
b)
Snhg26
Qiao et al., Nature Communications 2020
Expression (TPM)
Sperm
Oocyte
1-cell embryo
2-cell embryo
4-cell embryo
8-cell embryo
Blastocyst
c)
Day 5
n = 3
Non-targeting sgRNA
Snhg26 sgRNA
p=0.036
p=0.012
Mean fold change relative to non-targeting sgRNA (normalized to Rpl27)
Snhg26
Oct4
Nanog
Tfcp2l1
Pluripotency genes
d)
Snhg26
Qiao et al., Nature Communications 2020
Expression (x10^4 TPM)
Snhg26.217
Snhg26.216
e)
Wild-type
Snhg26KO + control
Snhg26KO + Snhg26.217
Snhg26KO + Snhg26.216
DAPI SOX2 GATA6
f)
Blastoid area (μm²)
g)
Blastoid cell number
h)
Proportion of SOX2+ cells (%)
Proportion of GATA6+ cells (%)
n.s.

Figure 9

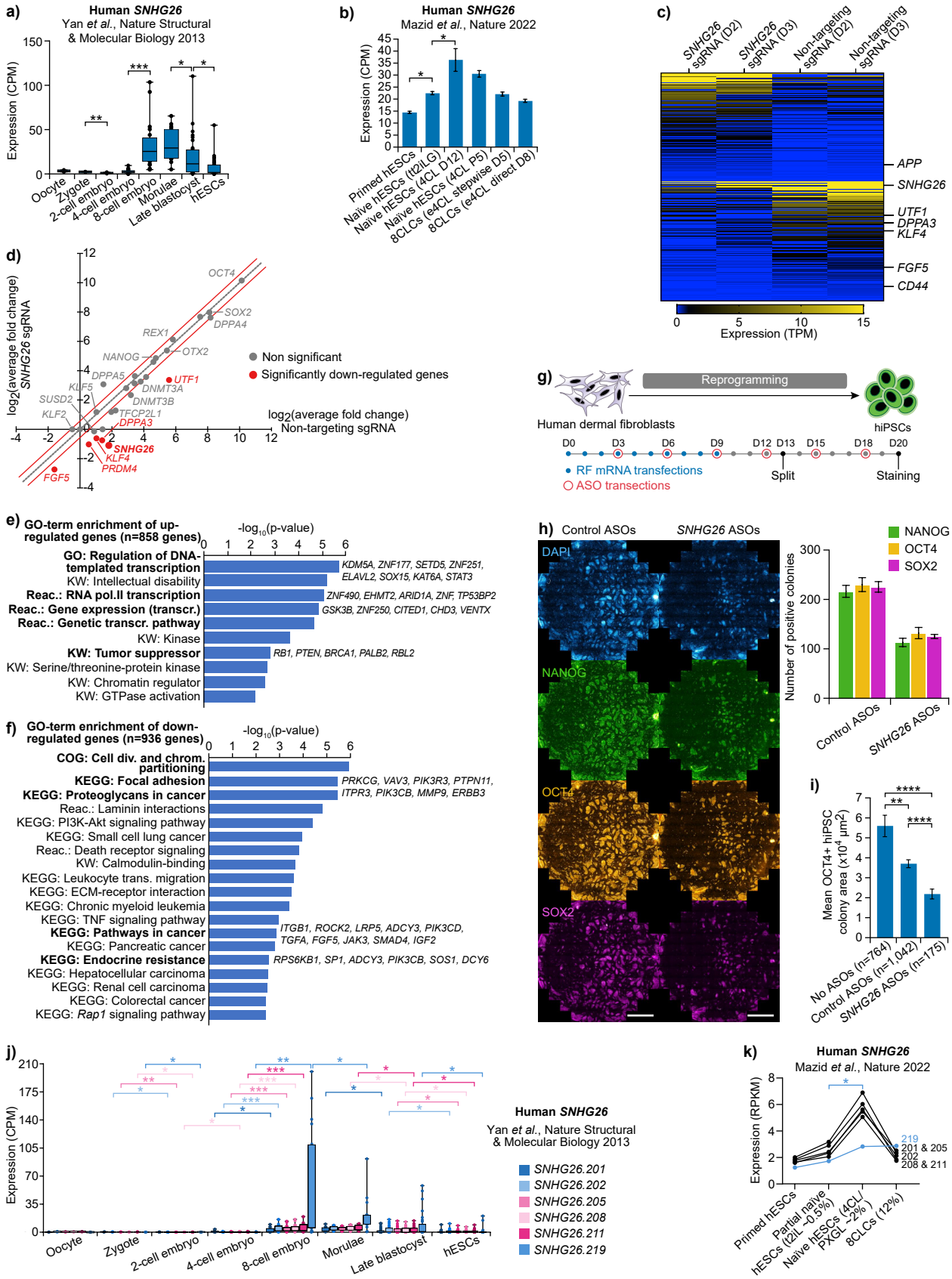

a)
Human SNHG26
Yan et al., Nature Structural & Molecular Biology 2013
Expression (CPM)
Oocyte
Zygote
2-cell embryo
4-cell embryo
8-cell embryo
Morulae
Late blastocyst
hESCs
b)
Human SNHG26
Mazid et al., Nature 2022
Expression (CPM)
Primed hESCs
Naïve hESCs (4CL D12)
Naïve hESCs (4CL P5)
Naïve hESCs (4CL stepwise D5)
8CLCs (e4CL direct D8)
c)
SNHG26 sgRNA (D2)
SNHG26 sgRNA (D3)
Non-targeting sgRNA (D2)
Non-targeting sgRNA (D3)
APP
SNHG26
UTF1
DPPA3
KLF4
FGF5
CD44
Expression (TPM)
d)
log2(average fold change) SNHG26 sgRNA
log2(average fold change) Non-targeting sgRNA
Non significant
Significantly down-regulated genes
OCT4
SOX2
DPPA4
REX1
OTX2
NANOG
UTF1
DPPA5
KLF5
DNMT3A
SUSD2
DNMT3B
KLF2
TFCP2L1
DPPA3
SNHG26
KLF4
PRDM4
FGF5
g)
Reprogramming
Human dermal fibroblasts
hiPSCs
D0
D3
D6
D9
D12
D13
D15
D18
D20
RF mRNA transfections
ASO transections
Split
Staining
e)
GO-term enrichment of up-regulated genes (n=858 genes)
-log10(p-value)
GO: Regulation of DNA-templated transcription
KW: Intellectual disability
Reac.: RNA pol.II transcription
Reac.: Gene expression (transcr.)
Reac.: Genetic transcr. pathway
KW: Kinase
KW: Tumor suppressor
KW: Serine/threonine-protein kinase
KW: Chromatin regulator
KW: GTPase activation
KDM5A, ZNF177, SETD5, ZNF251, ELAVL2, SOX15, KAT6A, STAT3
ZNF490, EHMT2, ARID1A, ZNF, TP53BP2
GSK3B, ZNF250, CITED1, CHD3, VENTX
RB1, PTEN, BRCA1, PALB2, RBL2
f)
GO-term enrichment of down-regulated genes (n=936 genes)
-log10(p-value)
COG: Cell div. and chrom. partitioning
KEGG: Focal adhesion
KEGG: Proteoglycans in cancer
Reac.: Laminin interactions
KEGG: PI3K-Akt signaling pathway
KEGG: Small cell lung cancer
Reac.: Death receptor signaling
KW: Calmodulin-binding
KEGG: Leukocyte trans. migration
KEGG: ECM-receptor interaction
KEGG: Chronic myeloid leukemia
KEGG: TNF signaling pathway
KEGG: Pathways in cancer
KEGG: Pancreatic cancer
KEGG: Endocrine resistance
KEGG: Hepatocellular carcinoma
KEGG: Renal cell carcinoma
KEGG: Colorectal cancer
KEGG: Rap1 signaling pathway
PRKCG, VAV3, PIK3R3, PTPN11, ITPR3, PIK3CB, MMP9, ERBB3
ITGB1, ROCK2, LRP5, ADCY3, PIK3CD, TGFA, FGF5, JAK3, SMAD4, IGF2
RPS6KB1, SP1, ADCY3, PIK3CB, SOS1, DCY6
h)
Control ASOs
SNHG26 ASOs
DAPI
NANOG
OCT4
SOX2
NANOG
OCT4
SOX2
Number of positive colonies
Control ASOs
SNHG26 ASOs
i)
Mean OCT4+ hiPSC colony area (x10^4 µm^2)
No ASOs (n=764)
Control ASOs (n=1,042)
SNHG26 ASOs (n=175)
j)
Expression (CPM)
Oocyte
Zygote
2-cell embryo
4-cell embryo
8-cell embryo
Morulae
Late blastocyst
hESCs
Human SNHG26
Yan et al., Nature Structural & Molecular Biology 2013
SNHG26.201
SNHG26.202
SNHG26.205
SNHG26.208
SNHG26.211
SNHG26.219
k)
Human SNHG26
Mazid et al., Nature 2022
Expression (RPKM)
219
201 & 205
202
208 & 211
Primed hESCs
Partial naïve hESCs (t2iL ~0.5%)
Naïve hESCs (4CL/PXGL ~2%)
8CLCs (12%)

Figure 10

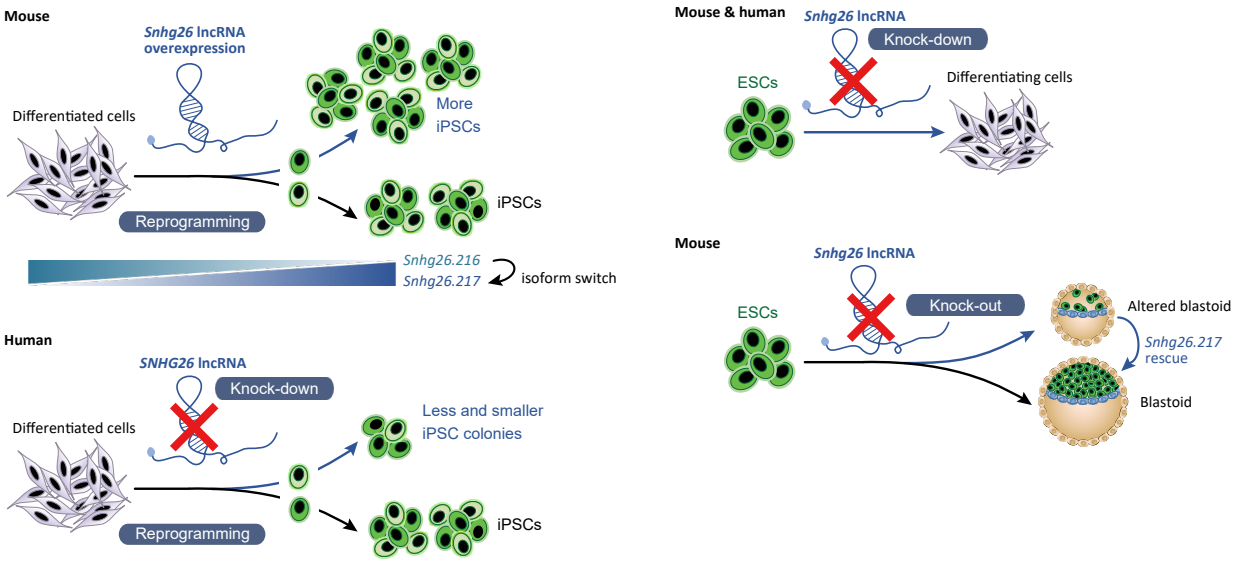

Mouse
Snhg26 lncRNA overexpression
Differentiated cells
More iPSCs
iPSCs
Reprogramming
Snhg26.216
Snhg26.217
isoform switch
Human
SNHG26 lncRNA
Knock-down
Differentiated cells
Less and smaller iPSC colonies
iPSCs
Reprogramming
Mouse & human
Snhg26 lncRNA
Knock-down
ESCs
Differentiating cells
Mouse
Snhg26 lncRNA
ESCs
Knock-out
Altered blastoid
Snhg26.217 rescue
Blastoid

Figure Legends

Figure 1: OKMS and OKS reprogramming systems that recapitulate expected gene expression

a) Schematic of the experimental reprogramming pipeline: NPCs harboring a *GFP* reporter under the control of the endogenous *Oct4* gene promoter were transfected with OKMS or OKS reprogramming factors driving the expression of an *mCherry* reporter. Co-transfection with a transposase, rtTA and a neomycin resistance gene allows for selection of cells with stably integrated cassettes through *piggyBac* transposition. Addition of doxycycline (DOX) to the medium induces overexpression of OKMS or OKS transgenes and leads to mCherry+ cells. Doxycycline is added either for the whole reprogramming process (upper track) or for six days only (lower track), leading to silencing of the transgenes and to GFP+ iPSCs. Samples were collected at different timepoints and processed for long-read RNA sequencing. ITR: Inverted Terminal Repeat (i.e. *piggyBac* transposition sequence); TRE: Tet Response Element; min. CMV: minimal CMV promoter; IRES: Internal Ribosome Entry Site; pA: poly(A) signal sequence; GOI: Gene Of Interest; rtTA: reverse tetracycline-controlled TransActivator; neoR: neomycin Resistance gene; NPCs: Neural Progenitor Cells; iPSCs: induced Pluripotent Stem Cells; RF: Reprogramming Factors (i.e. *Oct4* (O), *Klf4* (K), *c-Myc* (M) and *Sox2* (S)); endo.: endogenous; FBSL: FBS-LIF medium; 2iL: 2 inhibitors + LIF medium.

b) Plots of logged expression for immediate responder, intermediate reprogramming, early ESC-like and ESC-like genes (gene lists from Hussein *et al.*[16]), or mesenchymal-to-epithelial transition (MET) and epithelial-to-mesenchymal transition (EMT) genes in OKMS or OKS reprogramming systems. Lines represent average expression per timepoint and shaded areas represent the standard error of the mean (SEM). CPM: Counts Per Million reads; ESC: Embryonic Stem Cell.

c) Hierarchical clustering on top varying genes of the OKMS or OKS samples. The data is the same in OKMS and OKS for the NPC, FBSL and PSC group samples. repro.: reprogramming; PSC: Pluripotent Stem Cell.

d) Images of chimeric weaned mice obtained from aggregation of OKMS or OKS iPS clones with 8-cell CD-1 albino embryos for pluripotency assessment.

Figure 2: lrRNAseq reveals unprecedented isoform diversity and tight regulation of non-coding isoforms during reprogramming

a) Proportions of known and novel isoforms and of each isoform structural categories in the generated dataset after filtering out the transcripts lacking a poly(A) motif or with reverse-transcription (RT) switching events. FSM: Full-Splice Matches; ISM: Incomplete Splice Matches; NIC: Novel-In-Catalog; NNC: Novel-Not-in-Catalog.

b) Proportions of isoforms in each structural subcategory, according to SQANTI3 classification[54].TSS: Transcription Start Site; Comb.: Combination.

c) Proportions of coding isoforms going through or avoiding nonsense-mediated decay (NMD) among the different structural subcategories of the filtered dataset. Alt.: Alternative.

d) Proportions of coding and non-coding isoforms detected in the filtered dataset (pie chart), or separated by isoform structural category (bar graph), and box plots showing the coding probability of these isoforms. Values on the side of the box plots represent the median value of coding probabilities. Vertical dotted line represents the coding probability optimum cut-off indicating coding transcripts in mouse, according to the CPAT tool[128]. Interg.: Intergenic; Antis.: Antisense.

e) Mean percentages of coding and non-coding isoforms detected in the NPCs (n=8), during early (n=2), middle (n=4) or late (n=4 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) of the dataset, in total or **f)** separating annotated from novel isoforms. Error bars represent the standard deviation (STD) and each dot represents the replicates. Red outlines highlight significant data. P-values

were calculated comparing to NPCs using t-test with unequal variance. Non significant (n.s.) ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

g) Mean percentages of coding and non-coding NIC or NNC isoforms detected in NPCs (n=8), during early (n=2), middle (n=4) or late (n=4 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) of the dataset. Error bars represent the STD and each dot represents replicates. Red outlines highlight significant data. P-values were calculated comparing to NPCs using t-test with unequal variance. * < 0.05, ** < 0.01.

Figure 3: Alternative splicing follows a biphasic pattern during reprogramming

a) Heat map and plots depicting logged expression of pluripotency-related regulators of alternative splicing (gene lists from Cieply *et al.*[32] and Vivori *et al.*[35]) in OKMS or OKS reprogramming. Lines represent average expression and shaded areas represent SEM.

b) Schematic of the different types of local alternative splicing events analyzed in the filtered dataset, according to SUPPA2 classification[141].

c) Splicing events-based quantification of alternative splicing significantly more or less included, or **d)** isoform-based quantification of alternative splicing significantly more or less abundant, at timepoint transitions in OKMS or OKS reprogramming. Significance based on differential percent spliced-in (ΔPSI) ≥ 0.1 (more included) or ΔPSI ≤ -0.1 (less included) and an adjusted p-value ≤ 0.05. Horizontal dotted line represents the inflexion point in the biphasic alternative splicing pattern observed in reprogramming. AF: Alternative First exon; AL: Alternative Last exon; SE: Skipped Exon; AS: Alternative Splicing; MX: Mutually exclusive exons; RI: Retained Intron; A5': Alternative 5' splice site; A3': Alternative 3' splice site. Box plots represent the coding probability for significant alternatively spliced isoforms at timepoint transitions in OKMS or OKS reprogramming. Values to the right of box plots represent the average coding probabilities. Vertical dotted lines represent the coding probability optimum cut-off indicating

coding transcripts in mouse, according to the CPAT tool[128]. Black bars represent the medians. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, **** < 0.0001.

e) Distribution of 3'UTR lengths (log scale) in up-regulated isoforms from genes subject to alternative splicing during OKMS and OKS reprogramming phase transitions. Lines represent means of the NPCs to early datasets. Arrow points show change in distribution means in each transition. n.s. ≥ 0.05, * < 0.05, **** < 0.0001. UTR: UnTranslated Region; nt: nucleotides.

Figure 4: Selective interplay between gene expression and alternative splicing regulates reprogramming

a) Overlap between genes undergoing any significant alternative splicing detected from our lrRNAseq dataset (reprogramming from NPCs) and the datasets from Cieply *et al.*[32] (reprogramming from MEFs) and Vivori *et al.*[35] (reprogramming from B cells). MEF: Mouse Embryonic Fibroblast.

b) Schematic highlighting two alternative splicing events of exons 6 and 9 within the *Mbnl2* gene locus, and their respective consequences. Ex: Exon; aa: amino acids; N/C-ter: N-/C-terminus; ZNF: Zinc Finger domain; L: Linker domain; NLS: Nuclear Localization Signal; I: Interaction domain; LOF: Loss-Of-Function; PEST: PEST domain.

c) PSI values of *Mbnl2* exon 6 or exon 9 inclusion in phases of OKMS or OKS reprogramming. Lines represent average expression and shaded areas represent SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey dots).

d) Expression of *Mbnl2* gene in OKMS and OKS reprogramming phases. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01.

e) Logged p-values from GO-term enrichment analysis (biological processes and KEGG pathways) for the genes commonly spliced between our lrRNAseq dataset and the datasets from Cieply *et al.*[32] and Vivori *et al.*[35]. FDR ≤ 0.1.

f) Schematic highlighting the mutually exclusive alternative splicing of exons 7a and 7b within the *Fyn* gene locus, and their respective consequences. SH: Src Homology domain; U: Unique domain.

g) Expression of *Fyn* gene in OKMS and OKS reprogramming phases. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. ** < 0.01, *** < 0.001, **** < 0.0001.

h) PSI values of *Fyn* exon 7a inclusion in phases of OKMS or OKS reprogramming phases. Lines represent average expression and shaded areas represent SEM. P-values were calculated using t-test with unequal variance. ** < 0.01, **** < 0.0001. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey dots).

i) Mean expression of genes that negatively or positively correlate with *Fyn* in CRISPR screens from DepMap Public 25Q3 Release dataset[68], during the phases of OKMS and OKS reprogramming. Error bars represent the SEM. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey dots).

j) Percentages of isoforms with annotated PFAM domains or **k)** percentages of non-coding isoforms from protein-coding genes, for either all expressed genes, or genes that undergo alternative splicing, during the different phases of OKMS and OKS reprogramming. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles). * < 0.05, ** < 0.01. PCGs: Protein-Coding Genes.

Figure 5: Isoform switching has functional impacts on reprogramming

a) Number of up- or down-regulated isoforms between the different phases of OKMS and OKS reprogramming among the total isoforms (n=198,247) detected in the dataset. Significance based on log2(fold change) ≥ 2 for up-regulated isoforms or ≤ 0.5 for down-regulated isoforms with adjusted p-values ≤ 0.05. DEIs: Differentially Expressed Isoforms.

b) Schematic defining the difference between differential transcript usage (i.e. isoform switching) and differential isoform or gene expression. DTU/IS: Differential Transcript Usage/Isoform switch; DEG: Differentially Expressed Gene.

c) Logged p-values from GO-term enrichment analysis (biological processes) for genes undergoing isoform switch (IS) between D0 and the middle phase of OKMS or OKS reprogramming. FDR ≤ 0.1. Neg./Pos. reg.: Negative/Positive regulation.

d) Proportions of each gene biotype among the genes undergoing isoform switching, according to GENCODE annotation[144]. Bar graphs depict the observed percentage of protein-coding genes or lncRNA genes undergoing isoform switch, compared to their expected percentages found for genes undergoing differential isoform expression. P-values were calculated by Chi-squared test. **** < 0.0001. lncRNAs: long non-coding RNAs.

e) Overlap between the number of genes which undergo isoform switch between D0 and the middle phase of our reprogramming systems, based on either the ΔPSI events-based method or the differential isoform expression-based method.

f) Percentages of isoforms with annotated PFAM domains or non-coding isoforms from protein-coding genes, for genes that undergo isoform switching, during the different phases of OKMS and OKS reprogramming. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles). * < 0.05, ** < 0.01, **** < 0.0001.

g) Box plots of percentages of isoforms with annotated PFAM domains or the percentage of non-coding isoforms, for either all expressed genes, alternatively spliced genes (AS), or genes that undergo isoform switching, during the different phases of OKMS and OKS reprogramming. Central lines represent medians and boxes denote quartiles. * < 0.05, *** < 0.001, **** < 0.0001.

h) Schematic depicting the transcripts of the *Srsf2* gene detected in our dataset, as visualized with IGV software[143], and bar graph showing the coding probability of the *Srsf2* transcript categories. Horizontal dotted line represents the coding probability optimum cut-off indicating coding transcripts in mouse, according to the CPAT tool[128]. P-values were calculated using t-test with unequal variance. * < 0.05, **** < 0.0001. CDS: Coding Sequence.

i) Proportions of each *Srsf2* transcript category in OKMS and OKS reprogramming phases. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01. The data is the same in OKMS and OKS for NPCs to FBSL and ESCs to iPSCs comparisons because those samples were in common for OKMS and OKS.

Figure 6: Identification of a candidate non-coding transcript regulating the pluripotent state

a) Overlap between the number of lncRNA genes which undergo isoform switch between D0 and the middle phase of our reprogramming systems, based on either the ΔPSI events-based method or the differential isoform expression-based method.

b) Pipeline used to isolate lncRNA candidate genes potentially involved in the regulation of pluripotency induction.

c) Average expression of *2700038G22Rik* lncRNA at different timepoints of two 2° MEFs reprogramming RNAseq datasets[16,41] and one 2° TTFs reprogramming single-cell RNAseq dataset[18]. Error

bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01. TTFs: Tail-Tip Fibroblasts.

d) Expression of *2700038G22Rik* gene in OKMS and OKS reprogramming phases. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01.

e) Average expression of *SNHG26* lncRNA at different timepoints of a 2° human fibroblasts reprogramming RNAseq dataset[80]. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, **** < 0.0001.

f) Schematic depicting the transcripts of the *Snhg26* gene detected in our dataset and their level of support on the latest version of Ensembl database (release 113)[81]. TSL: Transcript Support Level.

g) Expression fold change normalized to NPCs, of two *Snhg26* lncRNA isoforms in OKMS and OKS reprogramming phases. Lines represent average expression and shaded areas represent % change. Pie charts represent the proportion of each transcript. P-values were calculated using t-test with unequal variance. * < 0.05, *** < 0.001. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles). TPM: Transcripts Per Million reads.

h) Proportions of cells expressing *Snhg26* isoforms in a 2° TTFs reprogramming single-cell RNAseq dataset[18].

i) Seurat gene module normalized expression scores for early and ESC-like genes (gene lists from Hussein *et al.*[16]) between cells expressing only isoform *Snhg26.216* or *Snhg26.217*. P-values were calculated by Mann-Whitney non-parametric U-test. n.s. > 0.05, * ≤ 0.05.

j) Seurat gene module normalized expression scores for early and ESC-like genes (gene lists from Hussein *et al.*[16]) between cells expressing coding isoforms from at least two of three markers of proper

reprogramming (*Dppa2, Utf1,* and *Esrrb*). P-values were calculated by Mann-Whitney non-parametric U-test. * < 0.05, ** < 0.01.

Figure 7: *Snhg26.217* isoform enhances the reprogramming route to pluripotency

a) Percentages of mCherry+ or **b)** GFP+ cells in FBSL + doxycycline medium, or after doxycycline removal in 2i-LIF medium, during reprogramming with OKS + control or OKS + *Snhg26.217* transgenes. Error bars represent SEM (n=3). P-values were calculated using t-test with paired samples. n.s. > 0.05, * ≤ 0.05. The data is the same as Supplementary Note 1b-c for OKS + control because experiments were conducted at the same time for OKMS, OKS and OKS + *Snhg26.217* lncRNA.

c) Schematic representing the possible effect of *Snhg26.217* expression on the dynamics of somatic cell reprogramming and on the results of a limiting dilution assay, depending on whether it increases cell proliferation or the kinetics of cells to reprogram.

d) Average number of mCherry+ cells and their mean doubling times from the OKS + control and OKS + *Snhg26.217* (n=3) reprogramming cell lines through the first 5 days of reprogramming. Error bars represent the SEM. P-values were calculated using t-test on paired samples. * < 0.05. Bar graph of mean cell doubling times between 24 and 120 hours of doxycycline induction. Error bars represent the STD. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05.

e) Average density of mCherry+ or **f)** GFP+ cells in FBSL + doxycycline medium, or after doxycycline removal in 2i-LIF medium, during reprogramming with OKS + control or OKS + *Snhg26.217* transgenes. Error bars represent the SEM (n=3). P-values were calculated using t-test with paired samples. n.s. > 0.05, * ≤ 0.05.

g) Logged density of GFP+ cells at the endpoint of reprogramming, using either transient transfections (n=5) or stable cell lines (n=5) to express the control or *Snhg26.217* transgenes. Error bars represent the

minimal and maximal values. Crosses represent averages, horizontal lines represent medians. P-values were calculated by Wilcoxon matched-pairs signed-rank test. * ≤ 0.05. The data is the same as each endpoint presented in Supplementary Fig. 13a-f.

h) Mean percentages of GFP-negative wells for the different cell densities at D12 of reprogramming in FBSL + doxycycline medium and at D27 of reprogramming in 2i-LIF medium (n=3), allowing the use of Poisson distribution. Error bars represent the SEM.

i) Average iPSC frequencies in OKS + control and OKS + *Snhg26.217* conditions at D12 and D27 (n=3) of reprogramming. Error bars represent the STD. P-values were calculated using t-test with unequal variance. * < 0.05.

j) Representative image of a chimeric mouse obtained from aggregation of OKS + *Snhg26.217* iPS clones with 8-cell CD-1 albino embryos for pluripotency assessment.

Figure 8: *Snhg26* is a developmental gene important for mouse pluripotency

a) Signal from ChIPseq data[16] at the genomic locus of *Snhg26*. Tracks represent 2º MEFs (D0) and their reprogrammed counterparts after 16 days of reprogramming (D16), 2º iPSCs and control ESCs.

b) Gene-level expression of *Snhg26* lncRNA in different stages of a mouse pre-implantation embryo dataset[13].

c) Mean fold changes in expression for *Snhg26* lncRNA and pluripotency-associated genes, assessed by RT-qPCR after *Snhg26* knock-down in mouse ESCs (n=3) using the CRISPRi system. Expression is normalized to a control sample with non-targeting single-guide RNA (sgRNA) and to a housekeeping gene. Error bars represent the STD. P-values were calculated using t-test with unequal variance. p: p-value.

d) Expression of *Snhg26* isoforms in different stages of a mouse pre-implantation embryo dataset[13].

e) Representative images in fluorescence microscopy of wild-type or *Snhg26*^{KO} blastoids obtained after 3 days of overexpression of either a control transcript or *Snhg26* isoforms. Scalebars = 20 µm. KO: Knock-Out.

f) Area of the wild-type (n=16), *Snhg26*^{KO} + control (n=11), *Snhg26*^{KO} + *Snhg26.217* (n=10), or *Snhg26*^{KO} + *Snhg26.216* (n=12) blastoids. Error bars represent the minimal and maximal values. Horizontal lines represent medians. P-values were calculated by ordinary one-way ANOVA with Tukey HSD *post hoc* test. * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

g) Total number of cells, or **h)** proportion of cells positive for SOX2 or GATA6 staining in the wild-type (n=9), *Snhg26*^{KO} + control (n=9), *Snhg26*^{KO} + *Snhg26.217* (n=10), or *Snhg26*^{KO} + *Snhg26.216* (n=10) blastoids. Error bars represent the minimal and maximal values. Horizontal lines represent medians. P-values were calculated by ordinary one-way ANOVA with Tukey HSD *post hoc* test. n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

Figure 9: *Snhg26* is a developmental gene important for human pluripotency

a) Average expression of *SNHG26* lncRNA in different stages of a human pre-implantation embryo dataset[92]. Dots represent outlier cells. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001. hESCs: human Embryonic Stem Cells.

b) Expression of *SNHG26* lncRNA in primed and *naïve* hESCs or 8CLCs samples from an available dataset[95]. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05. tt2iLG: titrated 2 inhibitors with LIF and Gö6983 medium; e4CL: enhance 4 chemical with LIF medium; 8CLCs: 8 Cell stage embryo-Like Cells.

c) Heat map depicting the expression values of genes significantly up- or down-regulated (n=1,794) at D2 and D3 of *SNHG26* knock-down in hESCs, compared to control hESCs with a non-targeting sgRNA.

Significance based on a Shannon entropy score below tenth percentile. Maximum expression value was set at ≥ 15 TPM.

d) Logged expression of pluripotency genes (n=29) (significantly decreased are shown in red) following *SNHG26* knock-down in hESCs, as compared to conditions with a non-targeting sgRNA. Significance based either on fold change ≤ 0.67 and p-value ≤ 0.05, or on a Shannon entropy score below tenth percentile.

e) Logged p-values from GO-term enrichment analysis for genes up-regulated or **f)** down-regulated following *SNHG26* knock-down in hESCs. GO-term significance based on fold enrichment ≥ 1.5, p-value ≤ 0.05 and FDR ≤ 0.1. KW: Key Words; Reac: Reactome pathway; pol.: polymerase; transcr.: transcription; COG: Clusters of Orthologous Genes; div.: division; chrom.: chromosome; trans.: transendothelial.

g) Schematic of *SNHG26* knock-down during human reprogramming. hiPSCs: human induced Pluripotent Stem Cells; RF: Reprogramming Factors; ASO: AntiSense Oligonucleotide.

h) Representative whole-well immunofluorescence images and quantification of iPSC colonies stained with DAPI, NANOG, OCT4, and SOX2 antibodies at D20 of human reprogramming with control or *SNHG26*-targeting antisense oligonucleotides transfection. Scalebars = 3,000 µm.

i) Mean area of OCT4-positive iPSC colonies at D20 of human reprogramming with control or *SNHG26*-targeting ASOs. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. ** < 0.01, **** < 0.0001.

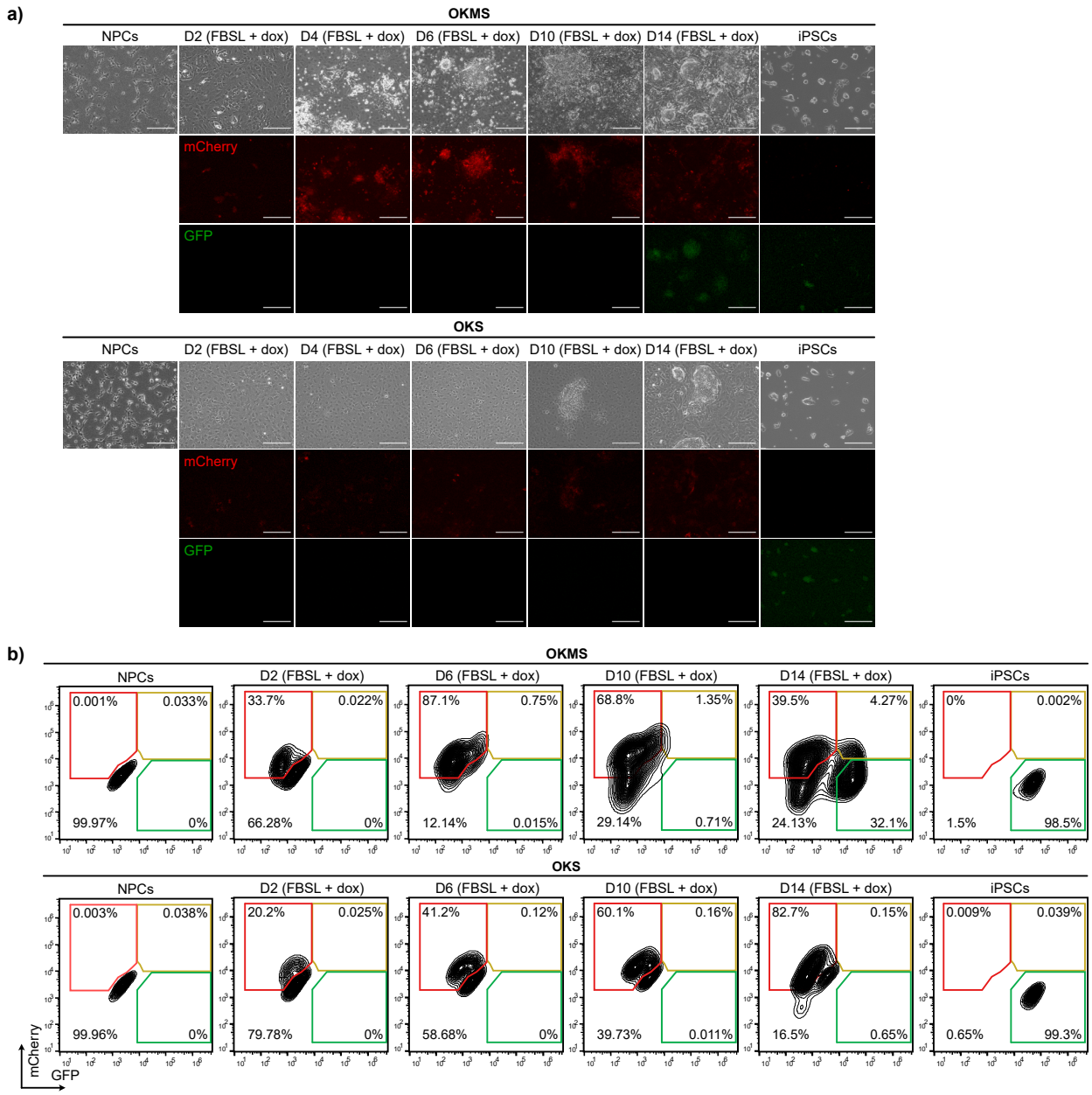
j) Expression of *SNHG26* isoforms in different stages of a human pre-implantation embryo dataset[92]. Dots represent outlier cells. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001.

k) Expression of *SNHG26* isoforms in primed and *naïve* hESCs or 8CLCs samples from a publicly available dataset[95]. P-values were calculated using t-test with unequal variance. * < 0.05. RPKM: Reads Per Kilobase of transcript per Million reads; PXGL: medium with PD0325901, XAV939, Gö6983 and LIF.

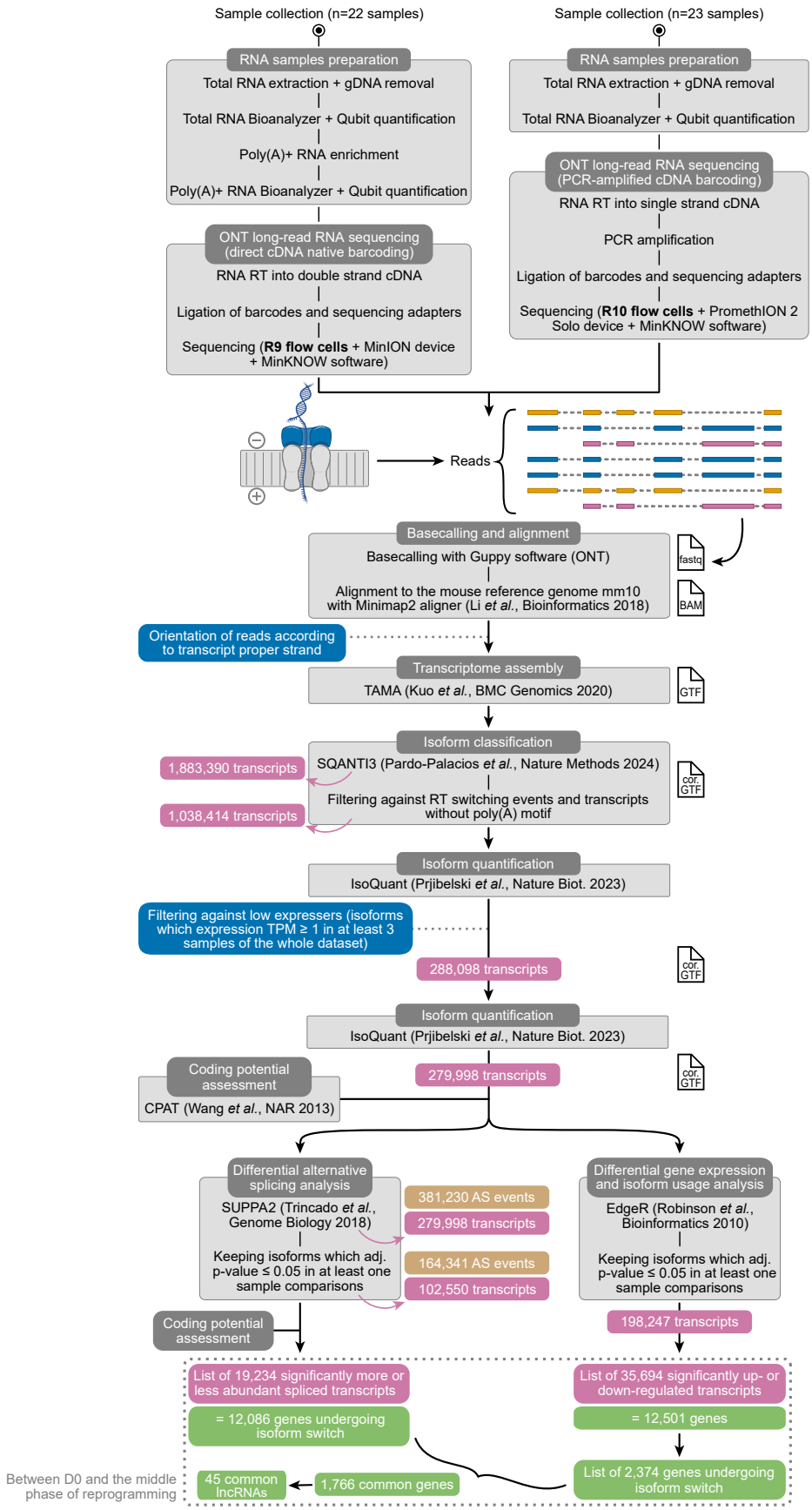
Figure 10: *Snhg26* is a conserved lncRNA that acts in the pluripotency regulatory network

Schematic representing the enhancing effect of *Snhg26* lncRNA on iPSC generation and the isoform switch that occurs during mouse reprogramming. Accordingly, knock-down of *SNHG26* during human reprogramming leads to less and smaller iPSC colonies. Comparatively, *Snhg26* knock-down promotes mouse and human ESC differentiation. *Snhg26* knock-out impairs mouse blastoids formation and generation of inner cell mass-like cells, which are both rescued by *Snhg26.217* isoform.

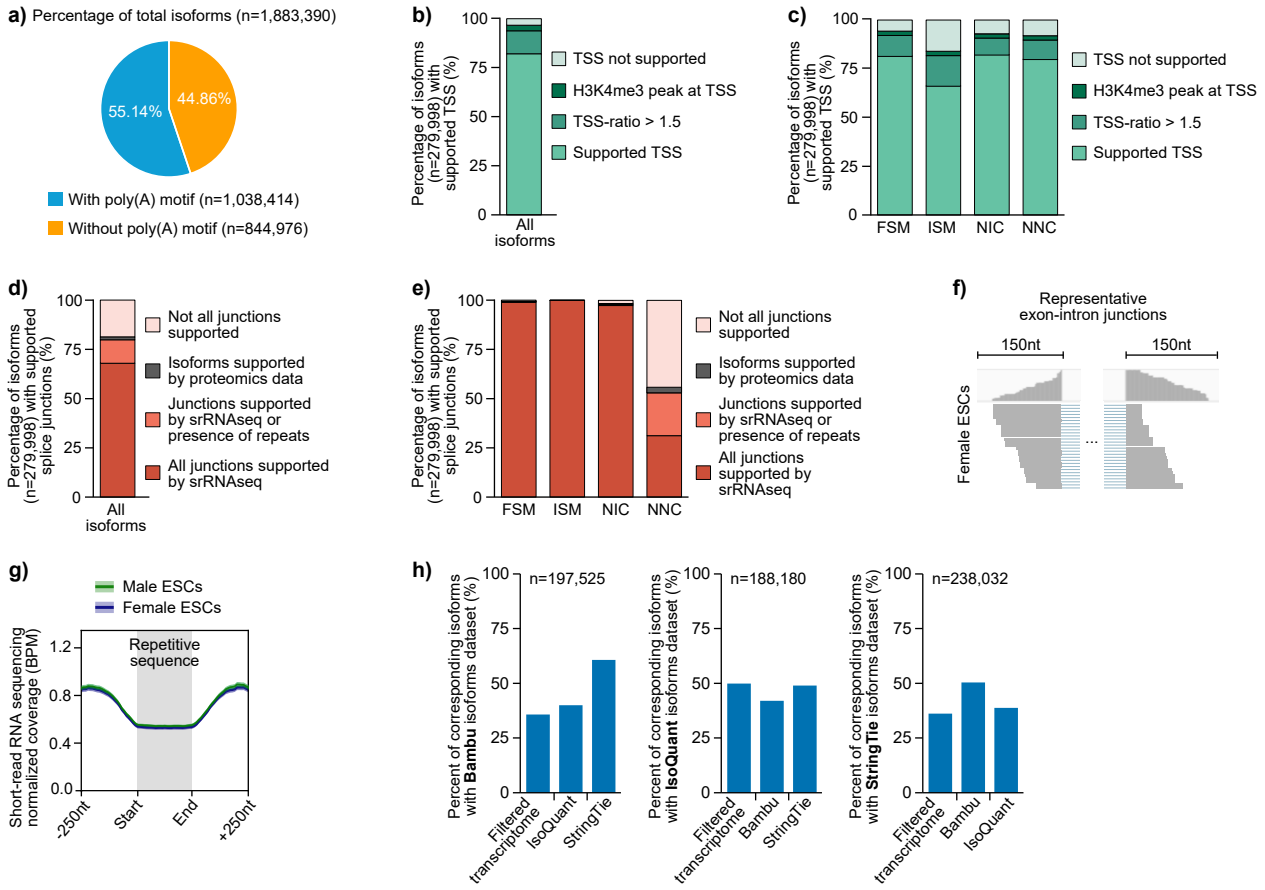
Supplementary Figure 1



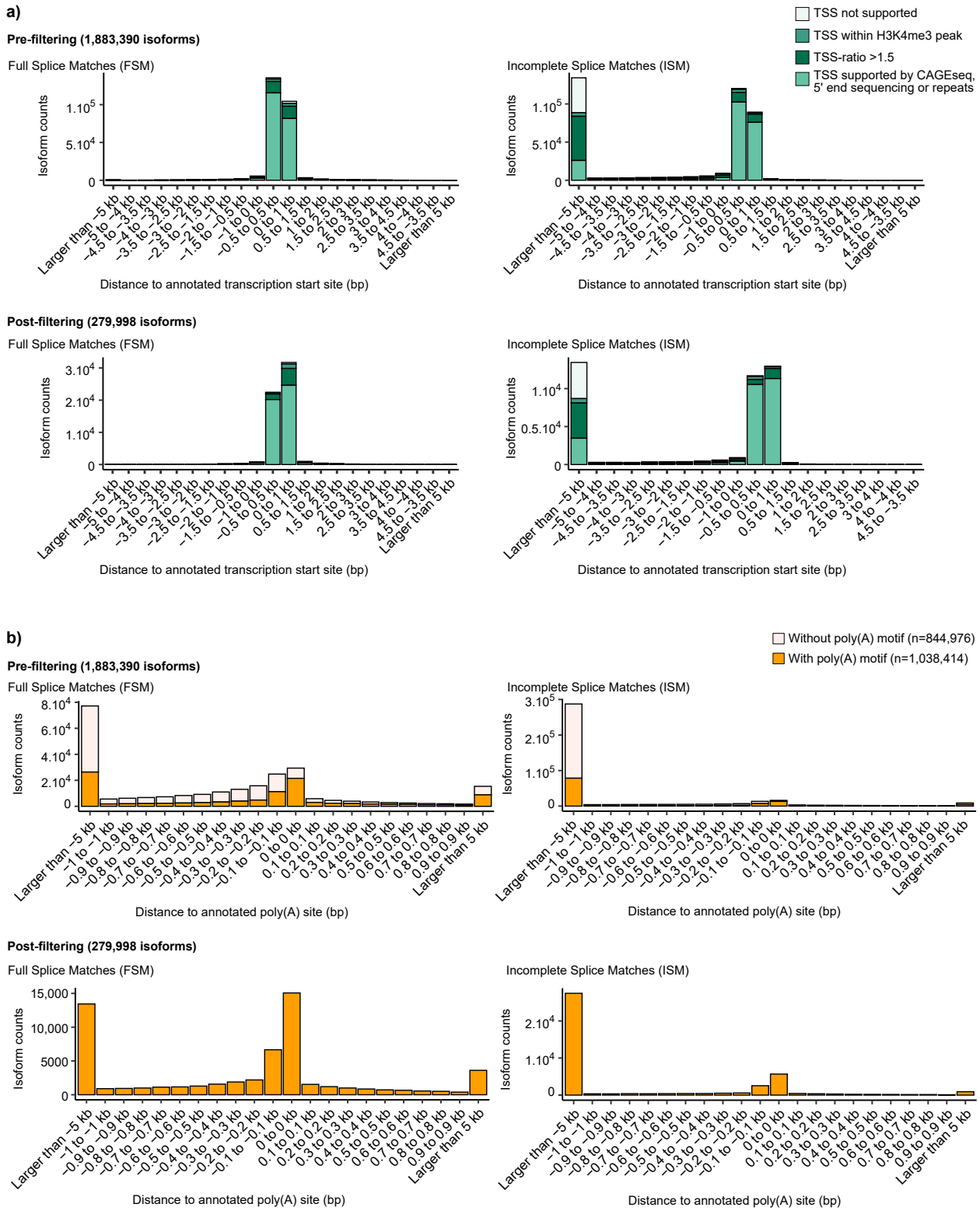
Supplementary Figure 2



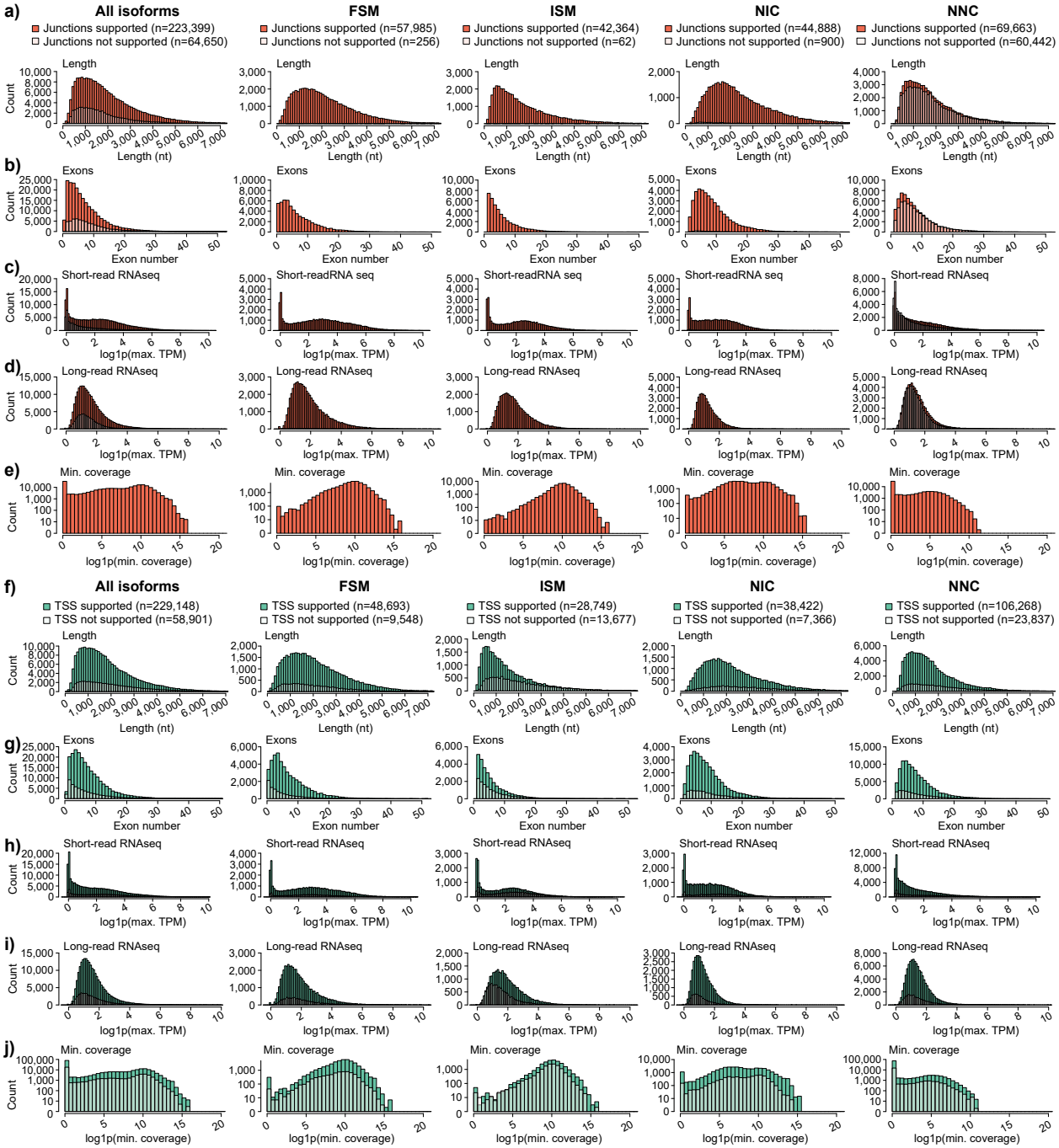
Supplementary Figure 3



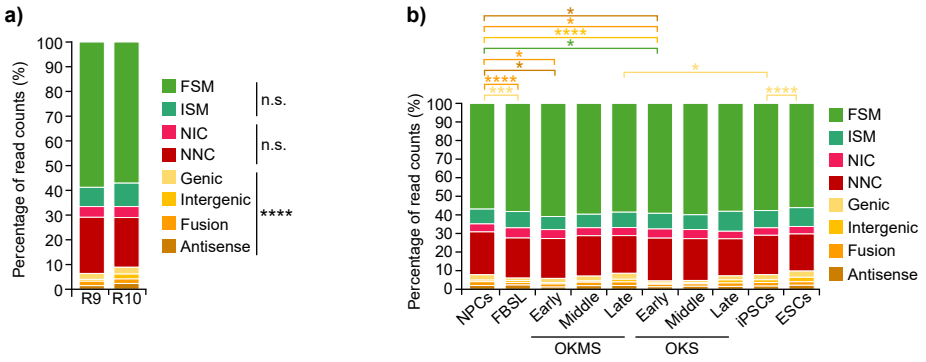
Supplementary Figure 4



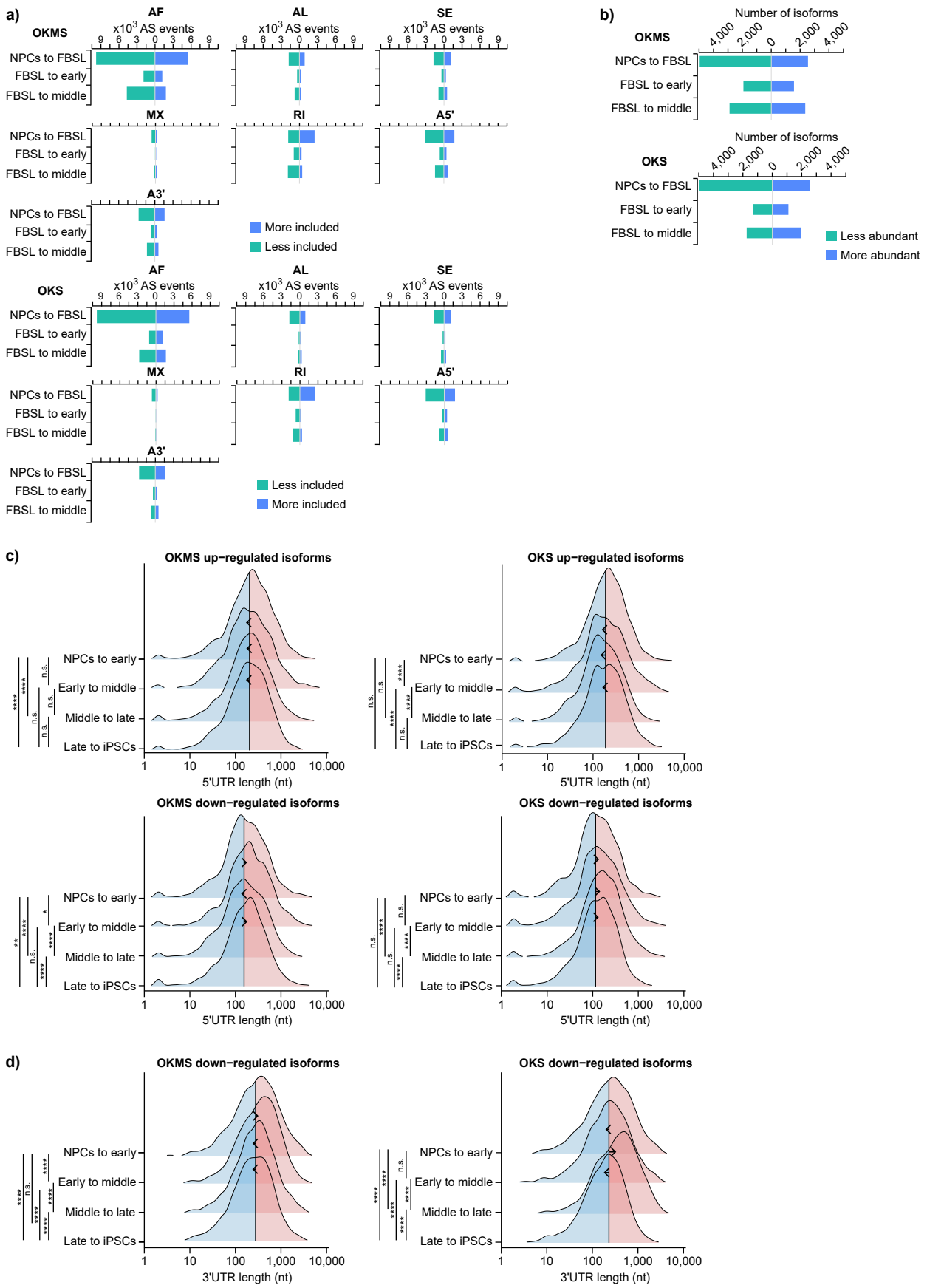
Supplementary Figure 5



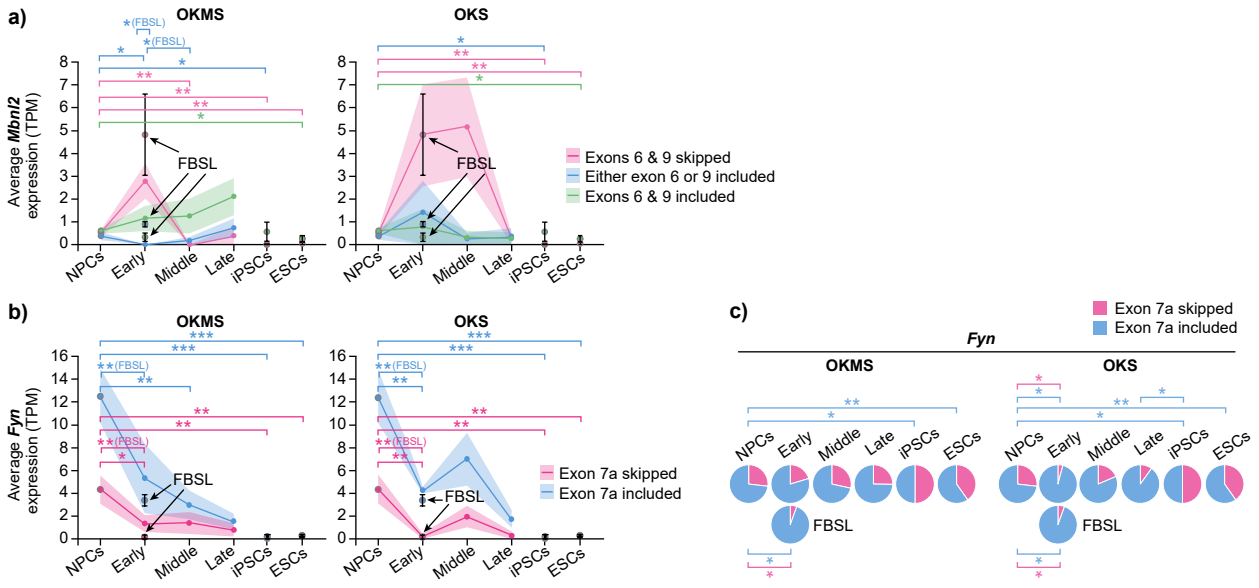
Supplementary Figure 6



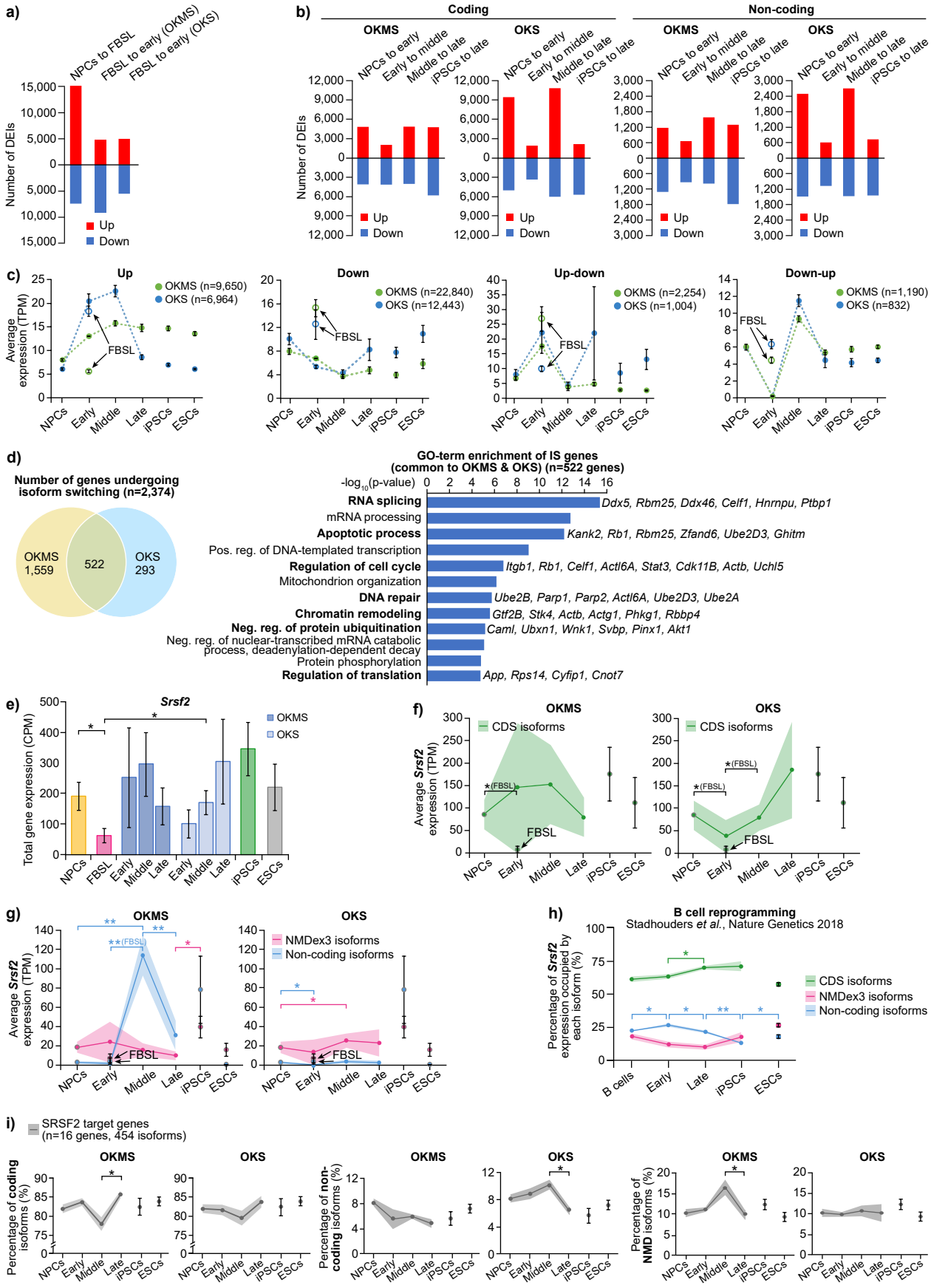
Supplementary Figure 7



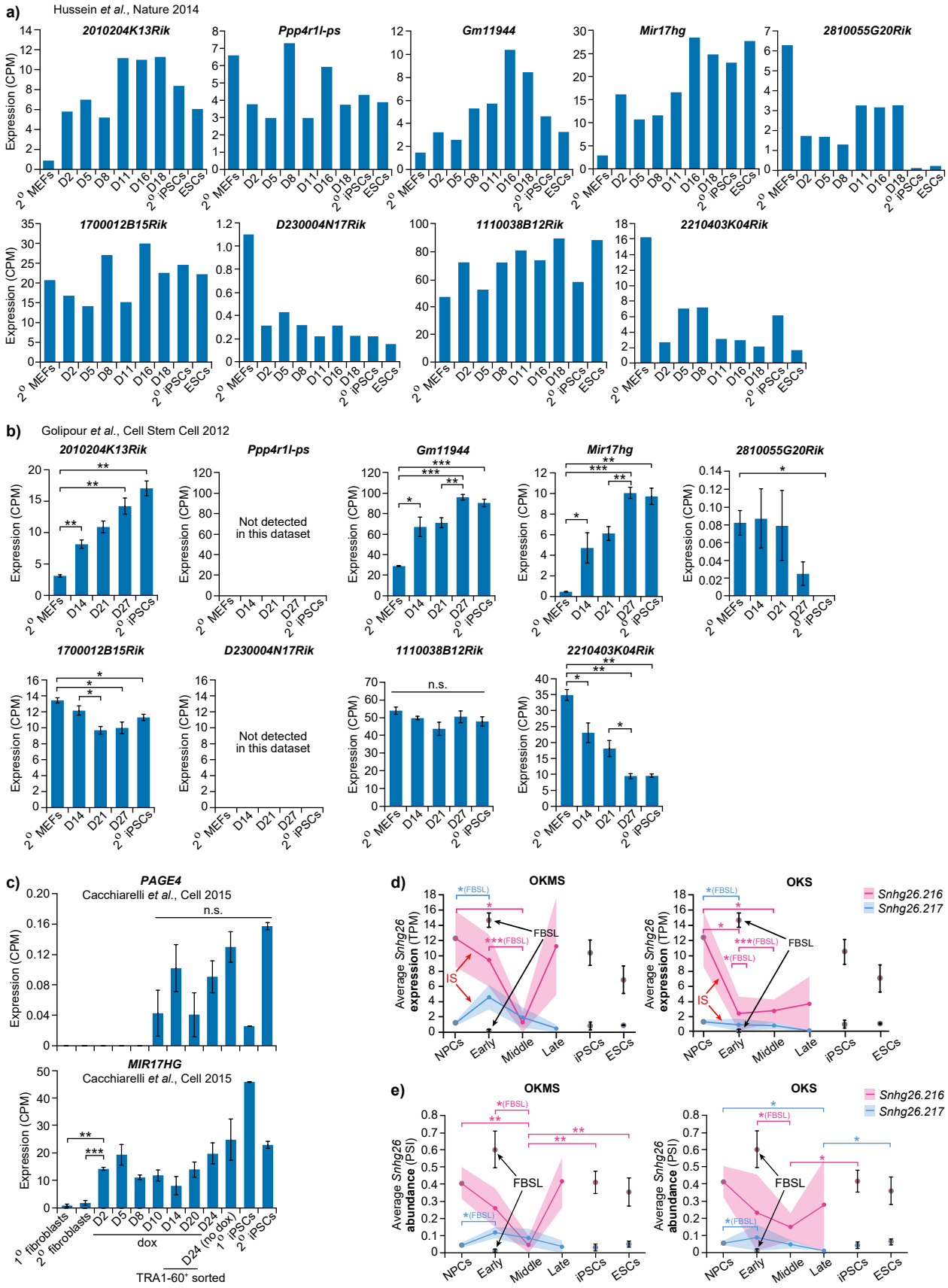
Supplementary Figure 8



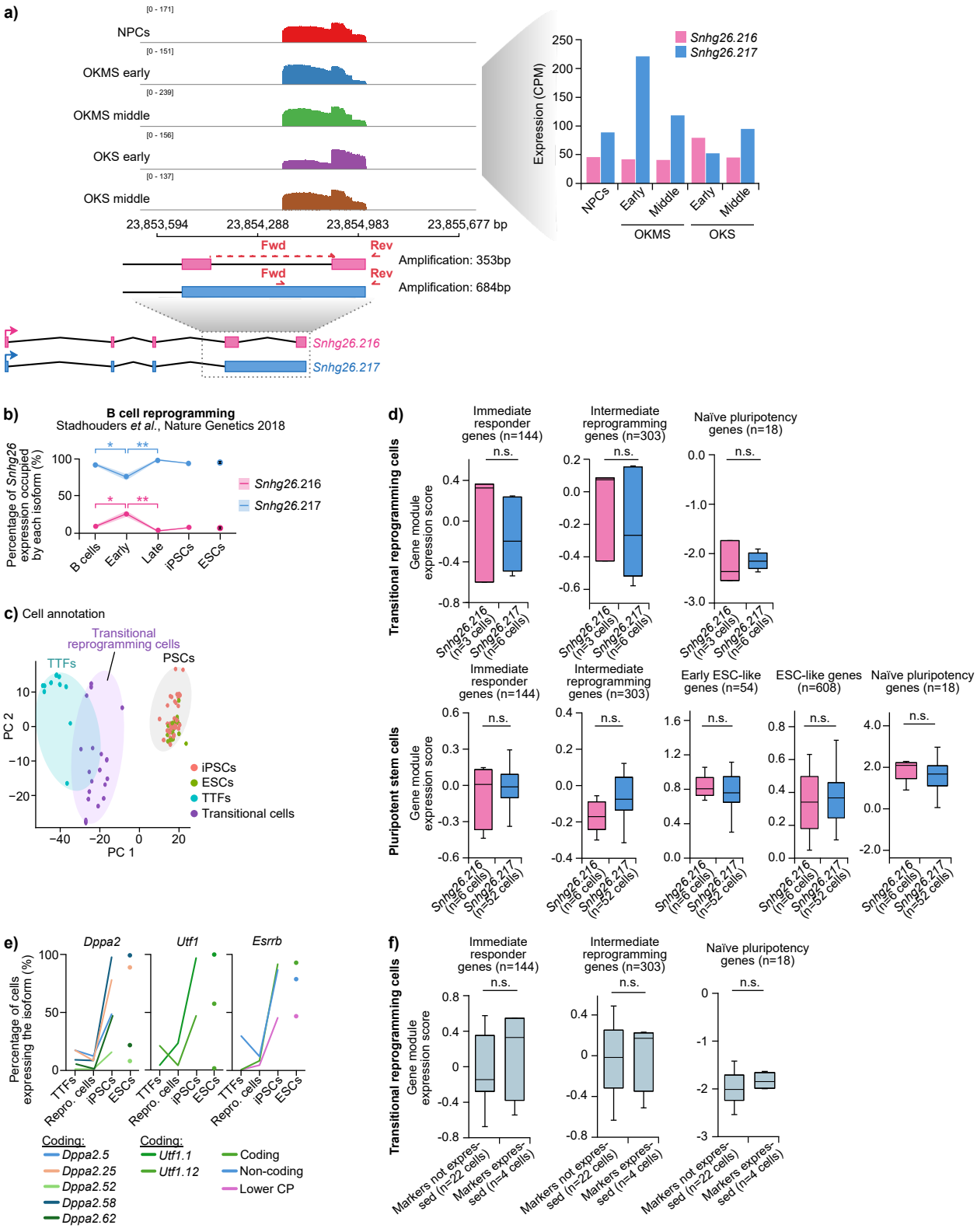
Supplementary Figure 9



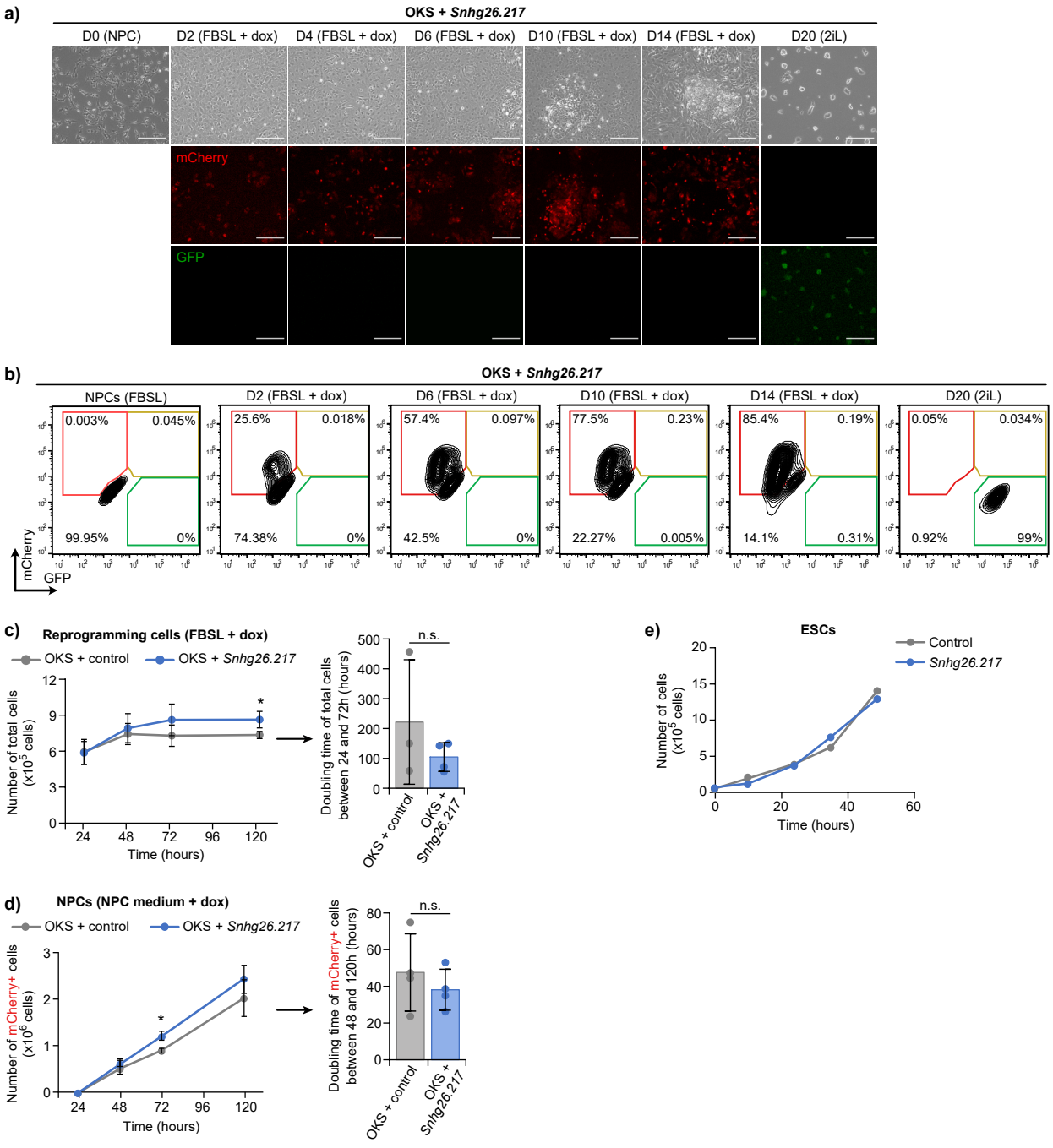
Supplementary Figure 10



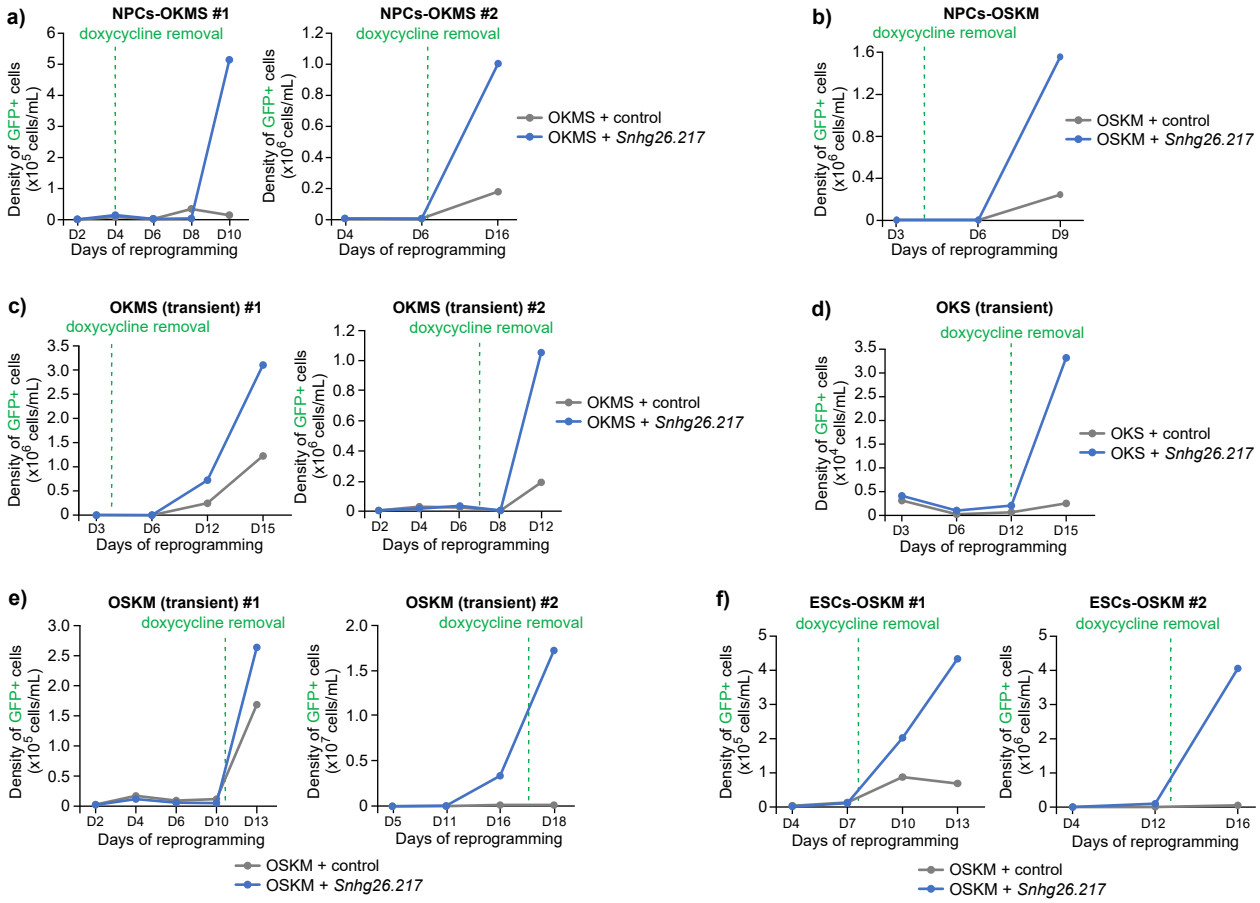
Supplementary Figure 11

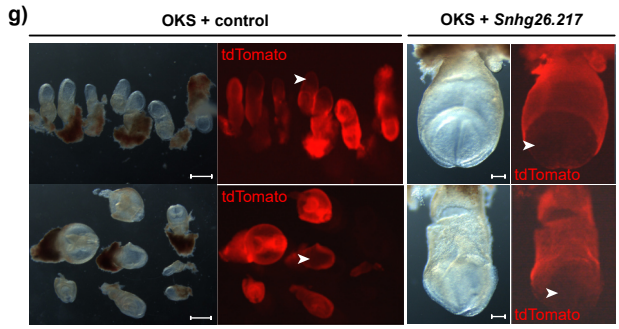


Supplementary Figure 12

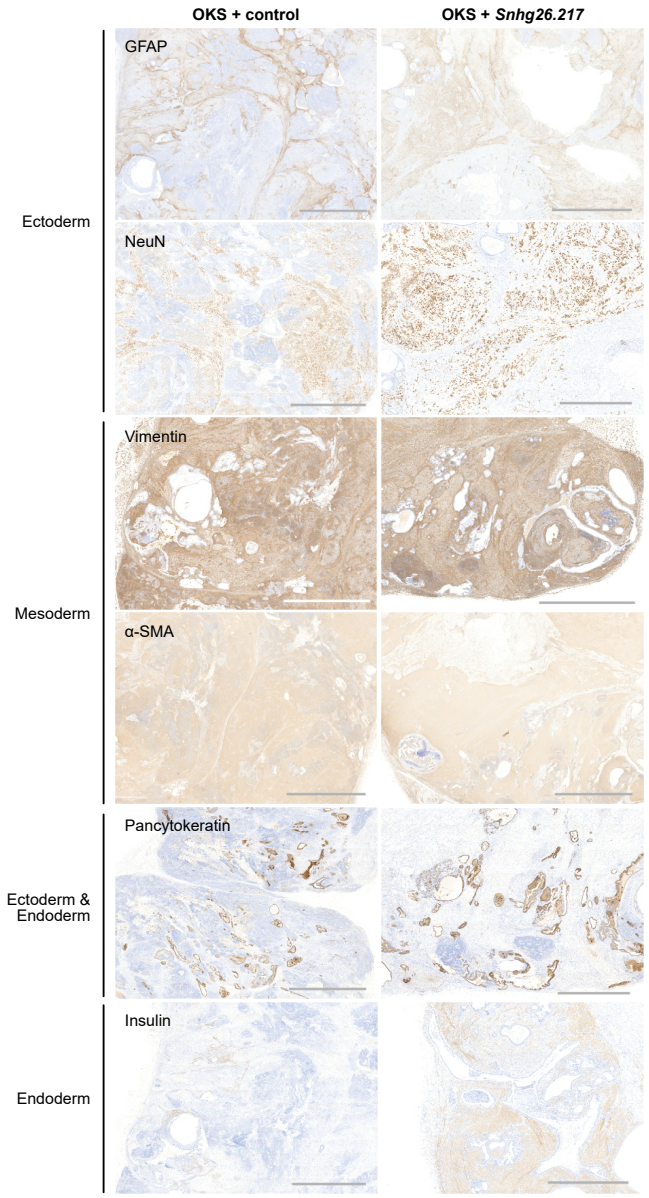


Supplementary Figure 13

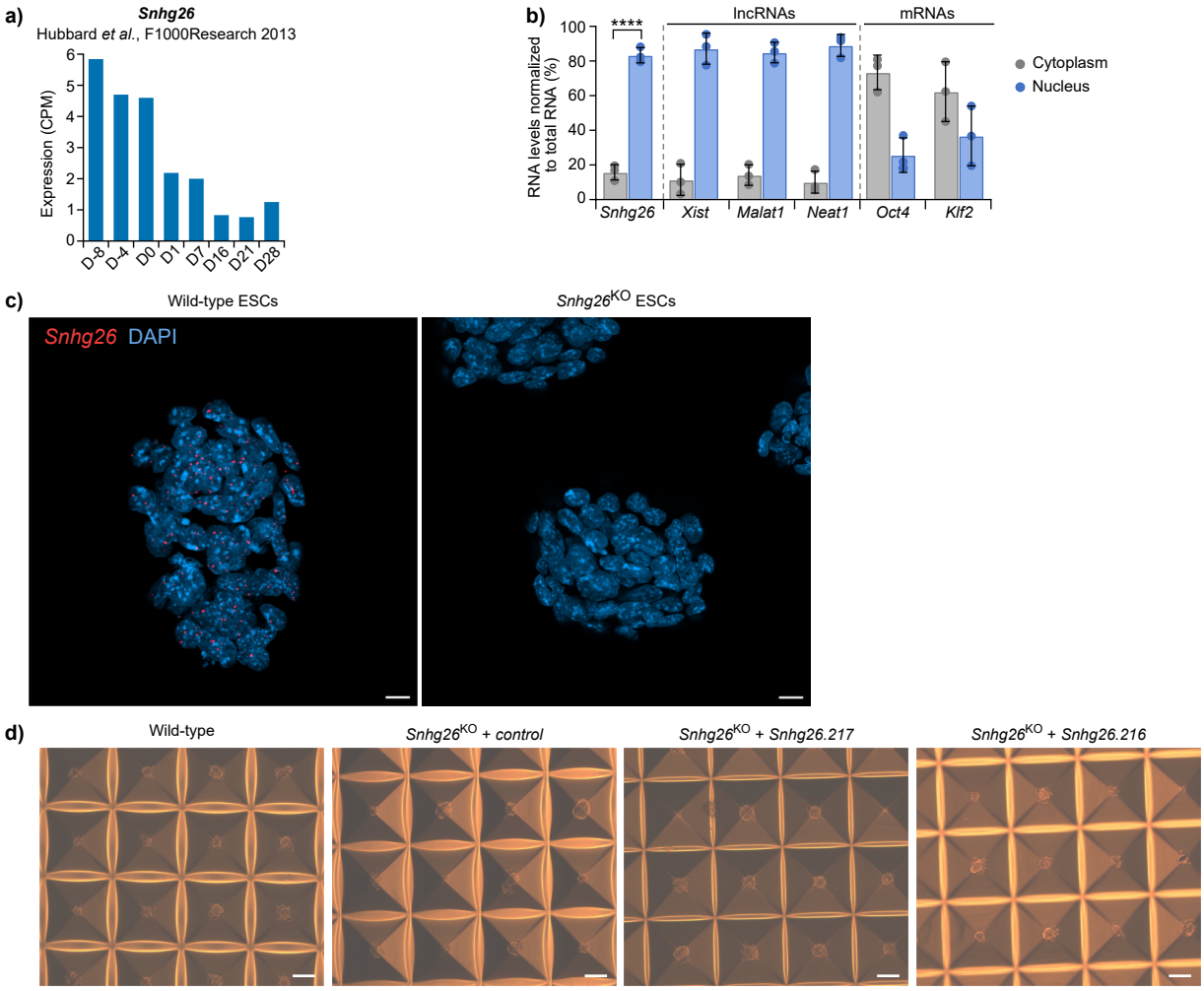




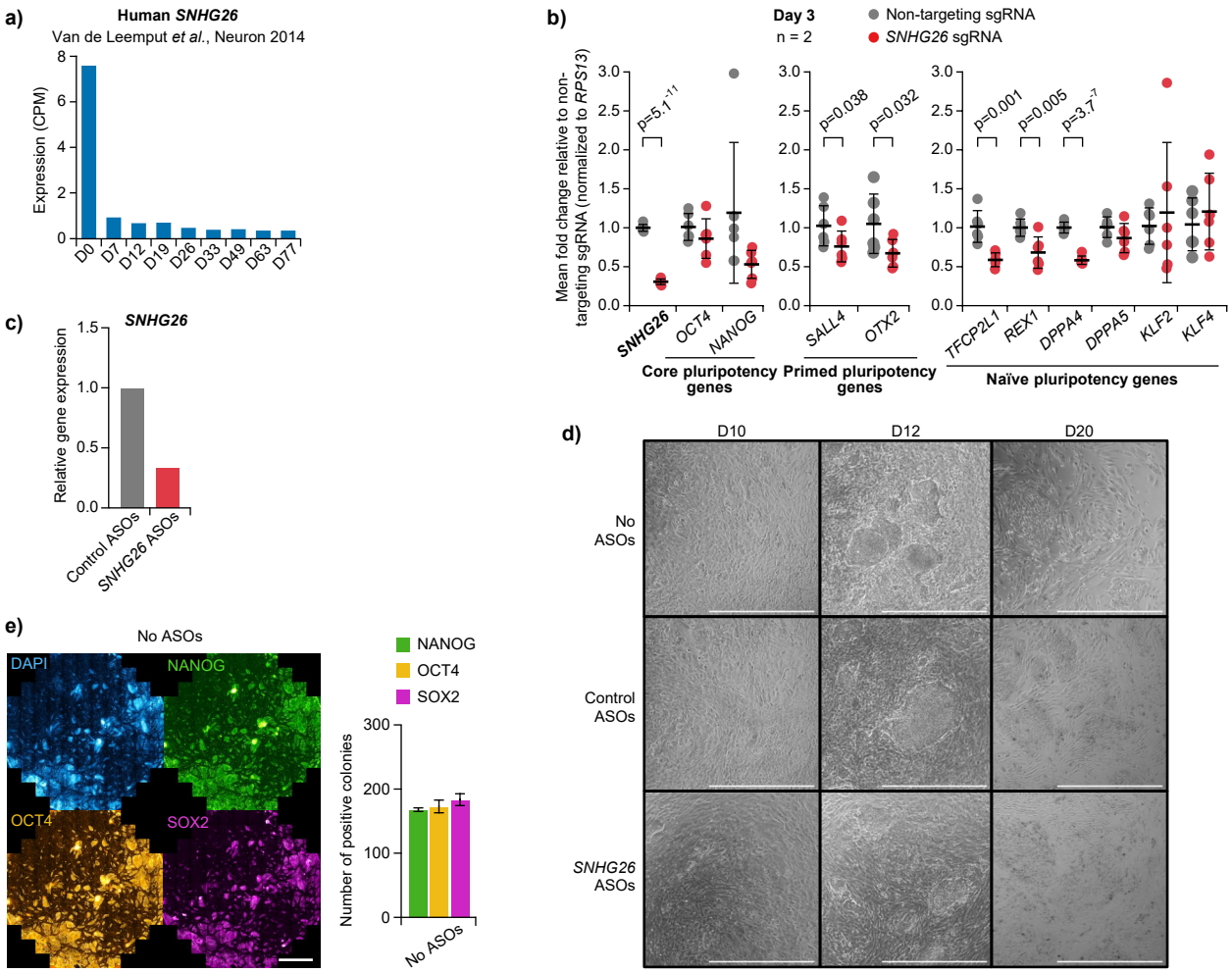
Supplementary Figure 14



Supplementary Figure 15



Supplementary Figure 16



Supplementary Figure 17

Supplementary Figure Legends

Supplementary Figure 1: *mCherry* and *GFP* reporter genes to follow reprogramming factors and endogenous *Oct4* expression

a) Representative images in brightfield or fluorescence microscopy of the NPCs, reprogramming cells at different timepoints, and final iPSCs (D20 2i-LIF) in the OKMS and OKS reprogramming systems. Scalebars = 50 µm.

b) Representative flow cytometry data allowing quantification of the percentages of reprogramming cells (mCherry+) and reprogrammed iPSCs (GFP+) throughout reprogramming with OKMS or OKS.

c) Number of up- or down-regulated genes between the early, middle and late phases of OKMS and OKS reprogramming. Significance based on FDR ≤ 0.05 and log2(fold change) ≥ 2 for up-regulated genes or ≤ 0.5 for down-regulated genes.

Supplementary Figure 2: Pipeline of the lrRNAseq dataset generation and analysis

Schematic overview of the experimental and analysis pipeline used to generate full-length transcript annotations of OKMS and OKS reprogramming samples. gDNA: genomic DNA; RT: Reverse Transcription; cDNA: complementary DNA; PCR: Polymerase Chain Reaction; ONT: Oxford Nanopore Technologies; cor. GTF: corrected GTF; adj. p-value: adjusted p-value.

Supplementary Figure 3: Filtering and validation of detected isoforms

a) Proportion of isoforms with or without a poly(A) motif among the total detected transcripts of the generated dataset.

b) Percentage of isoforms with transcription start sites supported by CAGEseq data or orthogonal approaches across all isoforms or **c)** within isoform structural categories. TSS-ratio: Ratio of read coverage downstream TSS to upstream TSS.

d) Percentage of isoforms with supported splice junctions across all isoforms or **e)** within isoform structural categories. srRNAseq: short-read RNA sequencing.

f) Representative view of short-read RNAseq split reads covering an intron on a genome browser. Boxes represent reads mapped to exons, lines represent read splitting over introns.

g) Short-read RNAseq coverage at and around exonic repetitive sequences in ESCs, normalized as bins-per-million (BPM). Lines represent average expression and shaded areas represent standard error.

h) Percentages of corresponding isoforms obtained by orthogonal transcriptome building tools (Bambu[50], IsoQuant[51] or StringTie[52]) when compared with the filtered transcriptome generated in this study and with each other.

Supplementary Figure 4: Distance from annotated isoform ends of the detected and filtered isoform datasets

a) Distance to annotated transcription start site and TSS support status, or **b)** distance to annotated poly(A) site and presence of poly(A) signal within isoforms for either the total transcriptome (available as processed data for GEO Series GSE282319), or the filtered transcriptome datasets. kb: kilobases; bp: base pair.

Supplementary Figure 5: Isoform quality metrics for supported and non-supported isoforms

a) to **e)** Isoform quality metrics for isoforms with or without splice junction support by short-read RNAseq, and **f)** to **j)** with or without CAGEseq support. Histograms represent **a)** and **f)** isoform lengths, **b)** and **g)** exon numbers, **c)** and **h)** expression in short-read RNAseq public datasets[129,138-140], **d)** and **i)** expression in our long-read RNAseq dataset, and **e)** and **j)** coverage at the lowest-covered splice junction (i.e. minimal (min.) coverage), shown with a log scale.

Supplementary Figure 6: Proportion of isoform categories during reprogramming

a) Proportions of isoform structural categories detected in the filtered dataset, separated by flow cell technology, or **b)** separated by reprogramming phase. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01, **** < 0.0001.

Supplementary Figure 7: Alternative splicing variations during reprogramming

a) Splicing events-based quantification of alternative splicing significantly more or less included, or **b)** isoform-based quantification of alternative splicing significantly more or less abundant, at each timepoint transition of OKMS or OKS reprogramming. Significance based on a ΔPSI ≥ 0.1 (more included) or ΔPSI ≤ -0.1 (less included) and an adjusted p-value ≤ 0.05. The data is the same in OKMS and OKS for NPCs to FBSL and ESCs to iPSCs comparisons because those samples were in common for OKMS and OKS.

c) Density plots showing the distribution of 5'UTR and **d)** 3'UTR lengths (log scale) in differentially expressed isoforms from genes subject to alternative splicing during OKMS and OKS reprogramming phase transitions. Lines represent means of the NPCs to early datasets. Arrow points show change in distribution means in each transition. n.s. ≥ 0.05, * < 0.05, ** < 0.01, **** < 0.0001.

Supplementary Figure 8: *Mbnl2* and *Fyn* isoforms expression

a) Expression of *Mbnl2* or **b)** *Fyn* isoforms in OKMS and OKS reprogramming phases. Lines represent the average expression value at each timepoint and the shaded area represents the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles).

c) Proportions of each *Fyn* transcript categories in OKMS and OKS reprogramming phases. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01. The data is the same in OKMS and OKS for NPCs to FBSL and ESCs to iPSCs comparisons because those samples were in common for OKMS and OKS.

Supplementary Figure 9: *Srsf2* isoforms switching

a) Number of up- or down-regulated isoforms between different phases of OKMS and OKS reprogramming among the total isoforms (n=198,247) detected in the dataset. Significance based on log2(fold change) ≥ 2 for up-regulated isoforms or ≤ 0.5 for down-regulated isoforms with p-values ≤ 0.05.

b) Number of coding and non-coding isoforms that are significantly up- or down-regulated between the different phases of OKMS and OKS reprogramming. Significance based on log2(fold change) ≥ 2 for up-regulated isoforms or ≤ 0.5 for down-regulated isoforms with p-values ≤ 0.05.

c) Average expression of isoforms differentially expressed between D0 and the middle phase of OKMS or OKS reprogramming. Error bars represent the SEM.

d) Overlap between the number of genes which undergo isoform switch between D0 and the middle phase of our reprogramming systems, and the corresponding GO-term enrichment analysis (biological processes) for the genes common to OKMS and OKS reprogramming. FDR ≤ 0.1.

e) Expression of *Srsf2* gene in OKMS and OKS reprogramming phases. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05.

f) and **g)** Expression of *Srsf2* isoforms in OKMS and OKS reprogramming phases. Lines represent the average expression value at each timepoint and the shaded area represents the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles).

h) Percentage of *Srsf2* expression occupied by its isoforms in a B cell reprogramming dataset[72]. Lines represent the average percentage value at each timepoint and the shaded area represents the SEM. * < 0.05, ** < 0.01.

i) Percentage of SRSF2 target genes (gene list from Han *et al.*[71]) expression coming from coding isoforms, non-coding isoforms or isoforms that undergo nonsense-mediated decay in OKMS and OKS phases. Lines represent average expression and shaded areas represent SEM. * < 0.05.

Supplementary Figure 10: Expression of candidate non-coding transcripts during reprogramming

a) and **b)** Average expression of the candidate lncRNAs other than *2700038G22Rik* at different timepoints of two 2° MEFs reprogramming RNAseq datasets[16,41]. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001.

c) Average expression of *PAGE4* and *MIR17HG* lncRNAs at different timepoints of a 2° human fibroblasts reprogramming RNAseq dataset[80]. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05, ** < 0.01, *** < 0.001.

d) Expression of two *Snhg26* lncRNA isoforms in OKMS and OKS reprogramming phases. Lines represent average expression and shaded areas represent SEM. P-values were calculated using t-test with

unequal variance. * < 0.05, *** < 0.001. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles).

e) PSI values of *Snhg26* transcripts abundance in the same samples of OKMS or OKS reprogramming. Lines represent the average PSI value at each timepoint and the shaded area represents the SEM. P-values were calculated using t-test with unequal variance. * < 0.05, ** < 0.01. The data is the same in OKMS and OKS for NPCs, FBSL, iPSCs and ESCs samples because those samples were in common for OKMS and OKS (grey circles).

Supplementary Figure 11: Validation and prediction of *Snhg26* isoforms' effect of the reprogramming outcome

a) Histograms representing the signal from long-read sequencing of the *Snhg26* isoforms validation PCR at the genomic loci of *Snhg26*. Tracks represent NPCs (D0) and samples from the early and middle phases of OKMS or OKS reprogramming. Localization of the primers specific to *Snhg26* isoforms and size of the corresponding amplicons are represented below. Quantifications of isoforms expression is represented in on the right. Fwd: Forward; Rev: Reverse.

b) Percentage of *Snhg26* expression occupied by its isoforms in a B cell reprogramming dataset[72]. Lines represent the average percentage value at each timepoint and the shaded area represents the SEM. * < 0.05, ** < 0.01.

c) Principal component analysis (PCA) for single-cell RNAseq[18] of reprogramming from TTFs to iPSCs. PC: Principal Component.

d) Seurat gene module normalized expression scores for different gene sets from Hussein *et al.*[16] between cells expressing only isoform *Snhg26.216* or *Snhg26.217,* for reprogramming transitional cells or pluripotent stem cells. P-values were calculated by Mann-Whitney non-parametric U-test. n.s. ≥ 0.05.

e) Percentage of cells expressing isoforms of markers of proper reprogramming (*Dppa2, Utf1,* and *Esrrb*) in a TTFs reprogramming dataset[18]. CP: Coding Potential.

f) Seurat gene module normalized expression scores for different gene sets from Hussein *et al.*[16] between cells expressing coding isoforms from at least two of three markers of proper reprogramming (*Dppa2, Utf1,* and *Esrrb*). P-values were calculated by Mann-Whitney non-parametric U-test. n.s. ≥ 0.05.

Supplementary Figure 12: *Snhg26.217* enhances reprogramming

a) Representative images in brightfield or fluorescence microscopy of the NPCs (D0), reprogramming cells at different timepoints, and final iPSCs (D20 2i-LIF) in the OKS + *Snhg26.217* reprogramming system. Scalebars = 50 µm.

b) Representative flow cytometry data allowing quantification of the percentages of reprogramming cells (mCherry+) and reprogrammed iPSCs (GFP+) throughout the reprogramming with OKS + *Snhg26.217*.

c) Average number of total cells from the OKS + control (n=3) and OKS + *Snhg26.217* (n=4) reprogramming cell lines in FBSL + doxycycline medium through the first 5 days of reprogramming. Error bars represent the SEM. P-values were calculated using t-test on paired samples. * < 0.05. Bar graph of mean cell doubling times between 24 and 72 hours of doxycycline induction. Error bars represent the STD. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05.

d) Average number of mCherry+ cells from the OKS + control (n=3) and OKS + *Snhg26.217* (n=4) reprogramming cell lines in NPC + doxycycline medium through the first 5 days of reprogramming. Error bars represent the SEM. P-values were calculated using t-test with unequal variance. * < 0.05. Bar graph of the mean cell doubling times between 48 and 120 hours of doxycycline induction. Error bars represent the STD and dot represents replicates. P-values were calculated using t-test with unequal variance. n.s. ≥ 0.05.

e) Number of cells from the ESC lines for the control (n=1) and *Snhg26.217* lncRNA (n=1) overexpression conditions in 2i-LIF + doxycycline medium through 2 days.

Supplementary Figure 13: *Snhg26.217* increases the number of iPSCs independently from the reprogramming system

a) Density of GFP+ cells in FBSL + doxycycline medium, or in 2i-LIF medium after doxycycline removal, during reprogramming with control or *Snhg26.217* transgenes using stable NPC lines with an OKMS (n=2) or **b)** OSKM (n=1) system.

c) Density of GFP+ cells in FBSL + doxycycline medium, or in 2i-LIF medium after doxycycline removal, during reprogramming with control or *Snhg26.217* transgenes using transient transfections with an OKMS (n=2), **d)** OKS (n=1), or **e)** OSKM (n=2) system.

f) Density of GFP+ cells in FBSL + doxycycline medium, or in 2i-LIF medium after doxycycline removal, during reprogramming with control or *Snhg26.217* transgenes using ESC line-derived NPCs with an OSKM (n=2) system.

g) Representative images in brightfield or fluorescence microscopy of E8-8.5 iPSC-embryo chimeras obtained from aggregation of OKS + control or OKS + *Snhg26.217* iPSC clones with 8-cell tdTomato-positive embryos for pluripotency assessment. Arrow points show examples of contribution from iPSCs. Scalebars = 200 µm (OKS + control); 100 µm (OKS + *Snhg26.217*).

h) Images of chimeric mice obtained from aggregation of OKS + *Snhg26.217* iPS clones with 8-cell CD-1 albino embryos, with different levels of contribution.

Supplementary Figure 14: iPSC clones generate teratomas

Representative images in brightfield of 1-month teratomas obtained from injection of OKS + control or OKS + *Snhg26.217* iPSC clones in mice and stained with germ layer differentiation markers for pluripotency assessment. Scalebars = 2,000 µm, except positive control for insulin scalebar = 1,000 µm.

Supplementary Figure 15: Characterization of *Snhg26* lncRNA

a) Expression of *Snhg26* lncRNA in an RNAseq dataset[87] of *in vitro* differentiation of mouse ESCs into cortical neurons.

b) Mean percentage of *Snhg26* lncRNA and other control mRNAs and lncRNAs in the nuclear or cytoplasmic compartments of mouse ESCs after fractionation (n=3), normalized to total RNA. Error bars represent the STD. P-values were calculated using t-test with unequal variance. **** < 0.0001.

c) Representative images of *Snhg26* lncRNA localization within mouse wild-type or *Snhg26*^{KO} ESCs in 2i-LIF medium, stained with DAPI and *Snhg26*-specific probes coupled to TSA Vivid dye 650. Scalebars = 10 µm.

d) Representative images in brightfield microscopy of the wild-type, *Snhg26*^{KO} + control, *Snhg26*^{KO} + *Snhg26.217*, or *Snhg26*^{KO} + *Snhg26.216* blastoids. Scalebars = 200 µm.

Supplementary Figure 16: Characterization of *SNHG26* knock-down in hESCs and during reprogramming

a) Expression of human *SNHG26* lncRNA in an RNAseq dataset[96] of *in vitro* differentiation of human ESCs into cortical neurons.

b) Mean fold changes in expression for human *SNHG26* lncRNA and pluripotency-associated genes, assessed by RT-qPCR after *SNHG26* knock-down in hESCs (n=2) using the CRISPRi system. Expression

is normalized to a control sample with non-targeting sgRNA and to a housekeeping gene. Error bars represent the STD from 6 technical replicates. P-values were calculated using t-test with unequal variance.

c) Fold change in expression for *SNHG26* lncRNA 12 days after first *SNHG26*-targeting ASOs transfection in human fibroblasts. Expression is normalized to a control sample with non-targeting ASOs.

d) Representative brightfield images of cells without transfection, transfection of control or *SNHG26*-targeting ASOs at D10, D12, and D20 of reprogramming. Scalebars = 1,000 µm.

e) Representative whole-well immunofluorescence images and quantification of iPSC colonies stained with DAPI, NANOG, OCT4, and SOX2 antibodies at D20 of reprogramming without ASO transfection. Scalebar = 3,000 µm.

Supplementary Figure 17: *SNHG26* isoforms

Schematic depicting the transcripts of *SNHG26* lncRNA on the Ensembl database (release 107)[104] and whether they are expressed in short-read and long-read RNAseq datasets of human pluripotent stem cells and early embryos[92,95,96,105-109]. Transcripts with similar exon-intron structure to mouse *Snhg26.216* or *Snhg26.217* isoforms are depicted in pink or in blue, respectively. Localization of ASOs used to knock-down *SNHG26* transcripts are shown in red.

References

1. McKinley, K.L., Castillo-Azofeifa, D. & Klein, O.D. Tools and Concepts for Interrogating and Defining Cellular Identity. *Cell Stem Cell* **26**, 632-656 (2020).
2. Rafelski & Theriot. Establishing a conceptual framework for holistic cell states and state transitions. *Cell* (2024).
3. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
4. He, S. et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* **21**, 294 (2020).
5. Lorenzi, L. et al. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat Biotechnol* **39**, 1453-1465 (2021).
6. Clough, E. et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res* **52**, D138-D144 (2024).
7. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).
8. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigo, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535-548 (2018).
9. Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* **21**, 597-614 (2020).
10. Oikonomopoulos, S. et al. Methodologies for Transcript Profiling Using Long-Read Technologies. *Frontiers in Genetics* **11**, 606 (2020).
11. Leung, S.K. et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports* **37**, 110022 (2021).
12. Torre, D. et al. Isoform-resolved transcriptome of the human preimplantation embryo. *Nature Communications* **14**, 6902 (2023).
13. Qiao, Y. et al. High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nature Communications* **11**, 2653 (2020).
14. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663-676 (2006).
15. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68-73 (2011).
16. Hussein, S.M.I. et al. Genome-wide characterization of the routes to pluripotency. *Nature* **516**, 198-206 (2014).
17. Clancy, J.L. et al. Small RNA changes en route to distinct cellular states of induced pluripotency. *Nature Communications* **5**, 5522 (2014).
18. Kim, Daniel H. et al. Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell* **16**, 88-101 (2015).
19. Lujan, E. et al. Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521**, 352-356 (2015).
20. Atlasi, Y. et al. The translational landscape of ground state pluripotency. *Nature Communications* **11**, 1617 (2020).
21. Dunn, S.J., Martello, G., Yordanov, B., Emmott, S. & Smith, A.G. Defining an essential transcription factor program for naïve pluripotency. *Science* **344**, 1156-1160 (2014).

22. Hainer, S.J., Bošković, A., McCannell, K.N., Rando, O.J. & Fazzio, T.G. Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319-1329.e11 (2019).
23. Panopoulos, A.D. et al. The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Res* **22**, 168-77 (2012).
24. Park, S.J. et al. Metabolome Profiling of Partial and Fully Reprogrammed Induced Pluripotent Stem Cells. *Stem Cells Dev* **26**, 734-742 (2017).
25. Guttman, M. et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).
26. Subramanyam, D. et al. Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nature Biotechnology* **29**, 443-448 (2011).
27. Miyamoto, T., Furusawa, C. & Kaneko, K. Pluripotency, Differentiation, and Reprogramming: A Gene Expression Dynamics Model with Epigenetic Feedback Regulation. *PLoS Comput Biol* **11**, e1004476 (2015).
28. Gabut, M. et al. An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell* **147**, 132-146 (2011).
29. Lu, Y. et al. Alternative Splicing of MBD2 Supports Self-Renewal in Human Pluripotent Stem Cells. *Cell Stem Cell* **15**, 92-101 (2014).
30. Dvir, S. et al. Uncovering the RNA-binding protein landscape in the pluripotency network of human embryonic stem cells. *Cell Reports* **35**, 109198 (2021).
31. Ohta, S., Nishida, E., Yamanaka, S. & Yamamoto, T. Global Splicing Pattern Reversion during Somatic Cell Reprogramming. *Cell Reports* **5**, 357-366 (2013).
32. Cieply, B. et al. Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. *Cell Reports* **15**, 247-255 (2016).
33. Toh, C.-Xu D. et al. RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. *Cell Reports* **15**, 2597-2607 (2016).
34. Kanitz, A., Syed, A.P., Kaji, K. & Zavolan, M. Conserved regulation of RNA processing in somatic cell reprogramming. *BMC Genomics* **20**, 100 (2019).
35. Vivori, C. et al. Dynamics of alternative splicing during somatic cell reprogramming reveals functions for RNA-binding proteins CPSF3, hnRNP UL1, and TIA1. *Genome Biology* **22**, 171 (2021).
36. Schlaeger, T.M. et al. A comparison of non-integrating reprogramming methods. *Nature Biotechnology* **33**, 58-63 (2015).
37. Hirsch, C.L. et al. Myc and SAGA rewire an alternative splicing network during early somatic cell reprogramming. *Genes & Development* **29**, 803-816 (2015).
38. Rand, T.A. et al. MYC Releases Early Reprogrammed Human Cells from Proliferation Pause via Retinoblastoma Protein Inhibition. *Cell Reports* **23**, 361-375 (2018).
39. Zviran, A. et al. Deterministic Somatic Cell Reprogramming Involves Continuous Transcriptional Changes Governed by Myc and Epigenetic-Driven Modules. *Cell Stem Cell* **24**, 328-341.e9 (2019).
40. Polo, Jose M. et al. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151**, 1617-1632 (2012).
41. Golipour, A. et al. A Late Transition in Somatic Cell Reprogramming Requires Regulators Distinct from the Pluripotency Network. *Cell Stem Cell* **11**, 769-782 (2012).

42. David, L. & Polo, J.M. Phases of reprogramming. *Stem Cell Research* **12**, 754-761 (2014).
43. Maherali, N. & Hochedlinger, K. Guidelines and techniques for the generation of induced pluripotent stem cells. *Cell Stem Cell* **3**, 595-605 (2008).
44. Samavarchi-Tehrani, P. et al. Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. *Cell Stem Cell* **7**, 64-77 (2010).
45. David, L. & Hirsch, C.L. How Tets and cytoskeleton dynamics MET in reprogramming. *Cell Stem Cell* **14**, 417-8 (2014).
46. Wernig, M., Meissner, A., Cassady, J.P. & Jaenisch, R. c-Myc Is Dispensable for Direct Reprogramming of Mouse Fibroblasts. *Cell Stem Cell* **2**, 10-12 (2008).
47. Nakagawa, M. et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnology* **26**, 101-106 (2008).
48. Buganim, Y. et al. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150**, 1209-1222 (2012).
49. Aich, M. et al. TOBF1 modulates mouse embryonic stem cell fate through regulating alternative splicing of pluripotency genes. *Cell Reports* **42**, 113177 (2023).
50. Chen, Y. et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**, 1187-1195 (2023).
51. Prjibelski, A.D. et al. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology* **41**, 915-918 (2023).
52. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-5 (2015).
53. Pardo-Palacios, F.J. et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Methods* **21**, 1349-1363 (2024).
54. Pardo-Palacios, F.J. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods* **21**, 793-797 (2024).
55. Buganim, Y., Faddah, D.A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics* **14**, 427-439 (2013).
56. Wong, J.J. & Schmitz, U. Intron retention: importance, challenges, and opportunities. *Trends Genet* **38**, 789-792 (2022).
57. Hangauer, M.J., Vaughn, I.W. & McManus, M.T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**, e1003569 (2013).
58. Ransohoff, J.D., Wei, Y. & Khavari, P.A. The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology* **19**, 143-157 (2018).
59. Guil, S. et al. Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat Struct Mol Biol* **19**, 664-70 (2012).
60. Barman, P., Reddy, D. & Bhaumik, S.R. Mechanisms of Antisense Transcription Initiation with Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis. *Noncoding RNA* **5**(2019).
61. Han, H. et al. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245 (2013).
62. Venables, J.P. et al. MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. *Nature Communications* **4**, 2480 (2013).

63. Shapiro, I.M. et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**, e1002218 (2011).
64. West, J.D. et al. The quantitative impact of 3'UTRs on gene expression. *Nucleic Acids Res* **53**(2025).
65. Nitschke, L., Hu, R.-C., Miller, A.N., Lucas, L. & Cooper, T.A. Alternative splicing mediates the compensatory upregulation of MBNL2 upon MBNL1 loss-of-function. *Nucleic Acids Research* **51**, 1245-1259 (2023).
66. Lee, J., Park, Y.-J. & Jung, H. Protein Kinases and Their Inhibitors in Pluripotent Stem Cell Fate Regulation. *Stem Cells International* **2019**, 1569740 (2019).
67. Brignatz, C. et al. Alternative Splicing Modulates Autoinhibition and SH3 Accessibility in the Src Kinase Fyn. *Molecular and Cellular Biology* **29**, 6438-6448 (2009).
68. 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).
69. Stadtfeld, M., Maherali, N., Breault, D.T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. *Cell Stem Cell* **2**, 230-240 (2008).
70. Smith, Z.D., Sindhu, C. & Meissner, A. Molecular features of cellular reprogramming and development. *Nature Reviews Molecular Cell Biology* **17**, 139-154 (2016).
71. Han, H. et al. Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. *Molecular Cell* **65**, 539-553.e7 (2017).
72. Stadhouders, R. et al. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **50**, 238-249 (2018).
73. Mu, P. et al. Genetic dissection of the miR-17~92 cluster of microRNAs in Myc-induced B-cell lymphomas. *Genes Dev* **23**, 2806-11 (2009).
74. Chakraborty, S., Mehtab, S., Patwardhan, A. & Krishnan, Y. Pri-miR-17-92a transcript folds into a tertiary structure and autoregulates its processing. *RNA* **18**, 1014-28 (2012).
75. Chen, C. et al. Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets. *Mamm Genome* **18**, 316-27 (2007).
76. Foshay, K.M. & Gallicano, G.I. miR-17 family miRNAs are expressed during early mammalian development and regulate stem cell differentiation. *Dev Biol* **326**, 431-43 (2009).
77. Gu, K.L. et al. Pluripotency-associated miR-290/302 family of microRNAs promote the dismantling of naive pluripotency. *Cell Res* **26**, 350-66 (2016).
78. He, X. et al. Human fibroblast reprogramming to pluripotent stem cells regulated by the miR19a/b-PTEN axis. *PLoS One* **9**, e95213 (2014).
79. Li, Z., Yang, C.S., Nakashima, K. & Rana, T.M. Small RNA-mediated regulation of iPS cell generation. *EMBO J* **30**, 823-34 (2011).
80. Cacchiarelli, D. et al. Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* **162**, 412-424 (2015).
81. Harrison, P.W. et al. Ensembl 2024. *Nucleic Acids Res* **52**, D891-D899 (2024).
82. Zhong, A., Harrison. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *The Journal of Immunology* (1996).

83. Ploemacher, R.E., van der Sluijs, J.P., Voerman, J.S. & Brons, N.H. An in vitro limiting-dilution assay of long-term repopulating hematopoietic stem cells in the mouse. *Blood* **74**, 2755-2763 (1989).
84. Sieburg, C., Müller-Sieburg. Limiting dilution analysis for estimating the frequency of hematopoietic stem cells: Uncertainty and significance. *Experimental Hematology* (2002).
85. Bernstein, B.E. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315-326 (2006).
86. Koche, R.P. et al. Reprogramming Factor Expression Initiates Widespread Targeted Chromatin Remodeling. *Cell Stem Cell* **8**, 96-105 (2011).
87. Hubbard, K.S., Gut, I.M., Lyman, M.E. & McNutt, P.M. Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. *F1000Res* **2**, 35 (2013).
88. Carlevaro-Fita, J. & Johnson, R. Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. *Molecular Cell* **73**, 869-883 (2019).
89. Mattick, J.S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*, 1-17 (2023).
90. Larson, M.H. et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8**, 2180-2196 (2013).
91. Gilbert, Luke A. et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442-451 (2013).
92. Yan, L. et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology* **20**, 1131-1139 (2013).
93. Davidson, K.C., Mason, E.A. & Pera, M.F. The pluripotent state in mouse and human. *Development* **142**, 3090-3099 (2015).
94. Messmer, T. et al. Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Rep* **26**, 815-824 e4 (2019).
95. Mazid, M.A. et al. Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315-324 (2022).
96. van de Leemput, J. et al. CORTECON: A Temporal Transcriptome Analysis of In Vitro Human Cerebral Cortex Development from Human Embryonic Stem Cells. *Neuron* **83**, 51-68 (2014).
97. Bao, Z., Zhang, W. & Dong, D. A potential prognostic lncRNA signature for predicting survival in patients with bladder urothelial carcinoma. *Oncotarget* **8**(2017).
98. Fan, Y. et al. Identification and validation of SNHG gene signature to predict malignant behaviors and therapeutic responses in glioblastoma. *Frontiers in Immunology* **13**, 986615 (2022).
99. Liu, M., Liu, C., Li, X. & Li, S. RP11-79H23.3 Inhibits the Proliferation and Metastasis of Non-small-cell Lung Cancer Through Promoting miR-29c. *Biochemical Genetics* **61**, 506-520 (2023).
100. Dong, C., Guo, Y., Wang, P., Yin, S. & Ge, X. Comprehensive analysis of disulfidptosis-related lncRNA features for prognosis and immune landscape prediction in colorectal cancer. *Frontiers in Oncology* **13**, 1287808 (2023).
101. Wu, Z.-H. et al. LncRNA SNHG26 promotes gastric cancer progression and metastasis by inducing c-Myc protein translation and an energy metabolism positive feedback loop. *Cell Death & Disease* **15**, 236 (2024).

102. Balachandran, S. & Narendran, A. The Developmental Origins of Cancer: A Review of the Genes Expressed in Embryonic Cells with Implications for Tumorigenesis. *Genes (Basel)* **14**(2023).
103. Li, D. et al. The lncRNA SNHG26 drives the inflammatory-to-proliferative state transition of keratinocyte progenitor cells during wound healing. *Nature Communications* **15**, 8637 (2024).
104. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res* **50**, D988-D995 (2022).
105. Chan, Y.-S. et al. Induction of a Human Pluripotent State with Distinct Regulatory Circuitry that Resembles Preimplantation Epiblast. *Cell Stem Cell* **13**, 663-675 (2013).
106. Au, K.F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences* **110**, E4821-E4830 (2013).
107. Takashima, Y. et al. Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell* **158**, 1254-1269 (2014).
108. Petropoulos, S. et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165**, 1012-1026 (2016).
109. Taubenschmid-Stowers, J. et al. 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell* **29**, 449-459.e6 (2022).
110. Loewer, S. et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics* **42**, 1113-1117 (2010).
111. Jiapaer, Z. et al. LincU Preserves Naïve Pluripotency by Restricting ERK Activity in Embryonic Stem Cells. *Stem Cell Reports* **11**, 395-409 (2018).
112. Haswell, J.R. et al. Genome-wide CRISPR interference screen identifies long non-coding RNA loci required for differentiation and pluripotency. *PLoS ONE* **16**, e0252848 (2021).
113. Yusa, K., Zhou, L., Li, M.A., Bradley, A. & Craig, N.L. A hyperactive piggyBac transposase for mammalian applications. *Proc Natl Acad Sci U S A* **108**, 1531-6 (2011).
114. Viswanathan, S. et al. Supplementation-dependent differences in the rates of embryonic stem cell self-renewal, differentiation, and apoptosis. *Biotechnology and Bioengineering* **84**, 505-517 (2003).
115. Mereau, A., Grey, L., Piquet-Pellorce, C. & Heath, J.K. Characterization of a binding protein for leukemia inhibitory factor localized in extracellular matrix. *The Journal of cell biology* **122**, 713-719 (1993).
116. Gertsenstein, M. et al. Efficient generation of germ line transmitting chimeras from C57BL/6N ES cells by aggregation with outbred host embryos. *PLoS One* **5**, e11260 (2010).
117. Gagnon, K.T., Li, L., Janowski, B.A. & Corey, D.R. Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading. *Nat Protoc* **9**, 2045-60 (2014).
118. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**, 148 (2016).
119. Schug, J. et al. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* **6**, R33 (2005).
120. Sherman, B.T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216-W221 (2022).
121. Mandal, P.K. & Rossi, D.J. Reprogramming human fibroblasts to pluripotency using modified mRNA. *Nat Protoc* **8**, 568-82 (2013).

122. Ran, F.A. et al. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281-2308 (2013).
123. Krentz, N.A.J., Nian, C. & Lynn, F.C. TALEN/CRISPR-Mediated eGFP Knock-In Add-On at the OCT4 Locus Does Not Impact Differentiation of Human Embryonic Stem Cells towards Endoderm. *PLoS ONE* **9**, e114275 (2014).
124. Jana, D. et al. Efficient self-organization of blastoids solely from mouse ESCs is facilitated by transient reactivation of 2C gene network. (2023).
125. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
126. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**(2023).
127. Kuo, R.I. et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**, 751 (2020).
128. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).
129. O'Malley, J. et al. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature* **499**, 88-91 (2013).
130. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
131. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
132. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
133. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-30 (2014).
134. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
135. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
136. Sasaki, K. et al. Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178-94 (2015).
137. Ohta, H. et al. In vitro expansion of mouse primordial germ cell-like cells recapitulates an epigenetic blank slate. *EMBO J* **36**, 1888-1907 (2017).
138. Abugessaisa, I. et al. refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J Mol Biol* **431**, 2407-2422 (2019).
139. Gu, X., Wang, M. & Zhang, X.O. TE-TSS: an integrated data resource of human and mouse transposable element (TE)-derived transcription start site (TSS). *Nucleic Acids Res* **52**, D322-D333 (2024).
140. Fort, A., Yamada, D., Hashimoto, K., Koseki, H. & Carninci, P. Nuclear transcriptome profiling of induced pluripotent stem cells and embryonic stem cells identify non-coding loci resistant to reprogramming. *Cell Cycle* **14**, 1148-55 (2015).
141. Trincado, J.L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40 (2018).

142. Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**, 169 (2015).
143. Robinson, J.T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-6 (2011).
144. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942-D949 (2023).