Long-read RNA sequencing identifies isoform switching of a novel long non-coding RNA that promotes somatic cell reprogramming

Victoire Fort[1,2,3], Gabriel Khelifi[1,2,3], Valérie Watters[1,2,3], Nobuko Yamanaka[4,5], Victoria Micha[1,2,3], Yojiro Yamanaka[4,5], Samer M.I. Hussein[1,2,3,*]

[1] Oncology division, CHU de Québec-Université Laval Research Centre; Quebec City, QC, Canada, G1J 5B3

[2] Faculty of Medicine, Université Laval, Quebec City, QC, Canada, G1V 0A6

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

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

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

* Correspondence to: Samer M. Hussein, Ph.D.
Department of Molecular Biology, Medical Biochemistry, and Pathology
Université Laval Cancer Research Centre
Oncology division - CRCHU de Québec – Université Laval
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

Key words

Long-read RNA sequencing, Isoform diversity, Long non-coding RNAs, Reprogramming, Pluripotency, Embryonic Stem Cells, Alternative splicing, Isoform switching

Abstract

The recent development of long-read RNA sequencing (lrRNAseq) technologies has paved the way to the exploration of RNA isoform diversity and its relevance in regulating cell fate and plasticity. However, drawing conclusions regarding the functionality of these newly identified isoforms is still very difficult. Here, we developed a very efficient system to reprogram mouse cells towards induced pluripotent cells and leveraged the strengths of lrRNAseq to study isoform diversity throughout reprogramming. We identified several isoforms derived from novel splice junctions occurring during this process. We demonstrate selective pressure on non-coding isoforms to be tightly regulated. We also show that isoform switching (IS) and diversification play a previously overlooked role in gene regulation to promote functional outcomes for reprogramming. Through IS analysis, we uncovered a novel long non-coding RNA (lncRNA) that undergoes IS during reprogramming to enhance the conversion of cells towards pluripotency. Knock-down of this lncRNA in mouse and human embryonic stem cells decreases the level of expression of pluripotency genes, suggesting that it is required for pluripotent stem cell maintenance. Together, our study provides a comprehensive resource to study full-length isoform usage and switching during reprogramming, but more importantly it demonstrates the power of lrRNAseq to identify functionally relevant gene isoforms in the context of pluripotency.

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 [1–3]. Despite the generation of a plethora of RNA sequencing (RNAseq) datasets by independent research teams and international consortiums[4–8], technical limitations have hindered our efforts to obtain a comprehensive and detailed overview of 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, RNAseq short reads lack continuity that can cover the full-length sequence of a transcript and thus don't have the power to unravel complex locus sequences and isoform diversity[9,10]. Long-read RNA sequencing (lrRNAseq) technologies, that are able to identify full-length transcripts[11–13], can overcome this issue to 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 cortexes, and highlighted changes in alternative splicing between human foetal and adult cortexes[14]. On the other hand,

studies in human[15] and mouse[16] 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[15]. In the mouse, Qiao *et al.* leveraged their lrRNAseq data analysis to identify new transcript isoforms where their depletion negatively impacted embryo development[16]. 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 further differentiation towards all the cell types of the body is crucial. Yet, the cellular pluripotent state is finely orchestrated by a complex regulatory network whose perturbations can lead to pathogenesis[17,18]. Understanding the regulatory network of PSCs enabled the seminal work by Yamanaka *et al.*, which led to the discovery that a small set of genes (*Oct4, Klf4, Sox2,* and *c-Myc*) can reprogram cells into induced pluripotent stem cells (iPSCs)[19,20]. Since then, extensive studies have been conducted to understand the principles driving this process of cellular plasticity[20,21] by characterizing reprogramming and PSCs in terms of epigenetics[22–24], transcriptomics[23,25–28], proteomics[29–31], transcription factor networks[32–34] and metabolomics[35,36]. 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)[20,32,37–39]. 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[23,28,40–47]. Indeed, short-read RNAseq analyses of mouse reprogramming 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[48,49]. Additionally, overexpressing or knocking-down splicing regulators enhanced or suppressed the efficiency of the reprogramming process, validating their importance in pluripotency regulation[26,48–55]. 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 the process of reprogramming mouse somatic cells to iPSCs using lrRNAseq, with the aim of identifying new gene isoforms regulating the pluripotent state. 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 selection during the mid-phases of reprogramming, and about 10% 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 one such lncRNA, *Snhg26*, enhances the reprogramming to the pluripotent state. 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, highlighting 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 genes, 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 pluripotency acquisition.

Results

1. Designing an efficient reprogramming system

Addressing isoform diversity in reprogramming is hindered by the low efficiency of the process[56,57]. Here, we designed two efficient 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, we devised our systems where *c-Myc* can be either included (OKMS) or omitted (OKS). This way, reprogramming-induced events that are independent of *c-Myc* can be isolated (please refer to **Supplementary Note** and **Supplementary Data 1** for more details). Briefly, we generated stable cell lines of mouse neural progenitor cells (NPCs), harboring a GFP transgene under the *Oct4* gene promoter, with doxycycline (DOX)-inducible reprogramming factor transgenes driving the expression of an *mCherry* reporter gene (**Fig. 1A** and **Extended Data Fig. 1A-B**). Upon DOX removal at D6, cells able to fully reprogram convert to iPSCs by naturally silencing OKMS or OKS transgenes and drive expression of *GFP* (**Fig. 1A** and **Extended Data Fig. 1A-B**).

To validate that our newly designed reprogramming systems faithfully recapitulate the known step-wise gene expression profiles of reprogramming, we performed whole transcriptome analysis using lrRNAseq at different timepoints of the process (**Fig. 1A** and **Table S1**). Analyzing expression of reprogramming and PSC marker gene sets as previously described[23], we found OKMS reprogramming to rapidly activate immediate responders, intermediate genes, and MET genes (genes associated with Mesenchymal-to-Epithelial transition). Early and late ESC-like genes were activated as early as D4 and D6, respectively (**Fig. 1B**). This was accompanied by decreased expression of EMT genes starting at D4 (**Extended Data Fig. 2A**) suggesting our system recapitulates MET, a hallmark process observed during

reprogramming to iPSCs[58,59]. 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 *Myc*, which is known to enhance reprogramming by increasing proliferation[60,61]. 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. Furthermore, similar to other studies[62,63], both OKMS and OKS systems showed a two-wave gene expression pattern, where gene expression is significantly affected both at the very beginning of reprogramming (upon DOX induction) and at the end (stabilization phase) (**Extended Data Fig. 2B**). Altogether, these results suggest that both our OKMS and OKS systems constitute very efficient and fast 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 global gene expression. This segregated the samples into 4 different categories: First the "NPC group" that gathers two replicates of wild-type NPCs, the OKMS-NPCs and the OKS-NPCs at D0. OKS reprogramming cells between D2 and D10 are found to be very similar to each other and to cluster with cells at D14 as well as OKMS cells between D2 and D6. Together they form a second group of "early and middle reprogramming phases group". Comparatively, OKS cells at D20 and OKMS cells between D10 and D20 are gathered in a separate "late reprogramming group" that resembles a "PSC group", composed of ESCs, OKMS- and OKS-iPSCs in 2i-LIF. This is consistent with the OKMS samples being quickly redirected towards the induced pluripotent state and achieving pluripotency reacquisition earlier than OKS samples (**Supplementary Data 1C**). Similar clustering performed independently on OKMS or OKS samples prompted us to group together samples belonging to close timepoints and to associate those groups to an "early", "middle" or "late" phase of the reprogramming process: D2 samples were assigned to the early phase, as they clustered closely to NPCs placed in FBS-LIF medium without reprogramming factors ("FBSL") (**Extended Data Fig. 2C**). D4 and D6, and D10 and D14 timepoints clustering together, they were chosen to constitute the "middle" and "late" phases of the OKMS reprogramming system, respectively. However, in the OKS system, D14 clustered separately from a group formed by D4, D6 and D10. Based on the previous analysis of differentially expressed genes, most changes seemed to happen only after the reprogramming cells reach D14 (**Extended Data Fig. 2B**), which prompted us to combine D14 and D20 together in a "late" phase for the OKS system (**Extended Data Fig. 2C**). These groups based on reprogramming phases were used later on for the rest of the analyses of our dataset. 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. Moreover, our gene expression analysis corroborates the idea that lrRNAseq is an adequate method to study gene expression.

2. LrRNAseq reveals unprecedented isoform diversity during reprogramming

In order to analyze isoform diversity and confidently identify functional isoforms in our newly designed reprogramming systems, we did the following: 1) To increase our replicate numbers, we sequenced 13 more samples in addition to our already sequenced 22 samples from above (**Extended Data Fig. 3A** and **Table S1**), 2) due to the rapid evolution of the technology, we used the latest R10 flow cells by Oxford Nanopore Technologies (ONT) and newer algorithms to obtain more accurate base calling, and 3) we focused on isoforms where a polyA motif can reliably be identified and that are not products of reverse transcription switching events to increase confidence in our detected isoforms (**Extended Data Fig. 3A** and **Table S1**). This led to the identification of 1,038,414 transcripts (55.14% of originally obtained transcripts) across the whole dataset and sequencing runs (**Extended Data Fig. 3B**). Of this overall filtered dataset, only 22.40% were annotated isoforms, whether they matched completely (10.51% "Full Splice Matches" or FSM) or partially (11.89% "Incomplete Splice Matches" or ISM) to the reference transcriptome (**Fig. 2A** and **Table S2**). This proportion is slightly lower than other lrRNAseq transcriptome analyses in the literature, but can easily be explained by the fact that those studies make use of older isoform annotation and analysis tools (e.g. TALON, in preprint[64], or of other lrRNAseq technologies (e.g. PacBio), and usually have lower depth or lower numbers of samples/less complex samples[14–16,65,66]. 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[21]. Thus, this could explain the identification of 805,861 (77.61%) novel isoforms. From the total, 69.84% of the transcripts were new isoforms from known genes, but with differing splice-junction combinations, with 13.75% coming from new combinations of known splice junctions ("Not-In-Catalog" or NIC) and 56.08% from novel splice junctions ("Novel-Not-in-Catalog" or NNC) (**Fig. 2A**). The remaining 7.77% of the detected isoforms were new transcripts generated from novel genes, with almost half of these (3.22%) originating from antisense genes, whereas intergenic (2.39%), genic (1.38%) and fusion (0.78%) isoforms accounted for the rest (**Fig. 2A**). Thus, our results are consistent with previous transcriptome analyses using lrRNAseq[14–16,66]. Moreover, investigating further subcategories of isoforms (**Fig. 2B**), NNC isoforms with at least one novel splice site constituted the main subcategory (50.94%), followed by isoforms with shorter 3' end (FSM *alternative 3' end* and ISM *5' fragment* subcategories, at 4.19% and 5.04%, respectively) and isoforms with *retained introns* (12.70% in total) (**Fig. 2C**). Interestingly, on average 48% of isoforms classified as having intron-retention were predicted to undergo nonsense-mediated decay (NMD) (NIC (48.83%), NNC (45.88%) and fusion (48.91%) *intron retention* subcategories) (**Fig. 2C**). This is consistent with the role of intron retention in NMD[67], but more importantly, recapitulates what we previously observed as a role for intron retention in regulating gene

expression during reprogramming[23]. Taken together, our results highlight the importance of lrRNAseq in identifying new isoforms and confirms 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[48,55].

We next analyzed for coding potential and found that the proportion of non-coding isoforms was significantly lower than coding isoforms (20.89% compared to 79.11%, respectively, **Extended Data Fig. 3C**). Furthermore, isoforms from known genes showed high coding probability (median between 0.69 and 0.80, **Fig. 2D** and **Table S2**) and coding potential (FSM: 83.63%; ISM: 82.40%; NIC: 90.46%, and NNC: 81.61%, **Extended Data Fig. 3C**). However, the opposite was true for novel genes. Intergenic and antisense isoforms showed a very low coding score (median of 0.11 for both, **Fig. 2D**), and more than 80% of these isoforms were non-coding (83.34% and 82.38%, respectively, **Extended Data Fig. 3C**), suggesting they are mainly lncRNAs. This is consistent with intergenic regions generally harboring lncRNAs[68,69] and with most antisense genes being transcribed as lncRNAs[70,71]. Moreover, about half (45.90%) of the new genic isoforms were non-coding as well (**Extended Data Fig. 3C**). These results suggest that the vast majority (about 70%) of new isoforms transcribed from novel genes detected during reprogramming are non-coding. Nonetheless, when it concerns novel isoforms from known genes, which constitute the majority of isoforms, there appears a selection pressure against non-coding isoforms. Altogether, these results demonstrate that the combination of our reprogramming systems with lrRNAseq identifies a diverse array of new isoforms during the conversion of differentiated cells to pluripotency, but whether these are functionally relevant or not remains to be determined.

3. Non-coding isoforms are exclusively regulated during reprogramming

To determine the dynamics of isoform diversity during reprogramming, more specifically how it affects non-coding isoforms, we first segregated isoform structural categories by reprogramming phase. The phases were selected based on the groupings described in **Extended Data Fig. 2D** and by merging replicates from our sequencing runs for increased stringency (**Extended Data Fig. 3A**, **Table S1** and **Table S2**). We detected significantly more isoforms from novel genes in iPSCs compared to NPCs in the corresponding structural categories, except for antisense isoforms (**Fig. 2E** and **Extended Data Fig. 3D**) regardless of coding potential (**Fig. 2F**). This increase occurred gradually during reprogramming, albeit at different phases in our two reprogramming systems (late phase for OKMS, and early phase for OKS) (**Fig. 2E** and **Extended Data Fig. 3D**). Interestingly, we observed that non-coding NNC isoforms were also affected in both reprogramming systems. The number of non-coding NNC isoforms dipped in the mid-phase of OKMS reprogramming and increased in early phase of OKS reprogramming but remained unchanged for coding NNC isoforms (**Fig. 2F**).

As NNC isoforms constitute the most abundant structural category of our dataset (more than 56%, **Fig. 2A**), and that 18.4% of these are non-coding (**Extended Data Fig. 3C**), non-coding NNC isoforms represent 107,097 transcripts. Together with non-coding intergenic and genic isoforms, they represent more than 134,390 transcripts, which accounts for about 13% of all detected isoforms. Altogether, these results suggest that novel non-coding transcripts, whether they were from known or novel genes, are tightly regulated and undergo selective pressure compared to coding isoforms during reprogramming.

4. 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[49]. Thus, we first assessed expression of 19 splicing regulators known to modulate alternative splicing in pluripotent stem cells and reprogramming[49,55], 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 **Table S3**). 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[43,48,49,55,72–74]. 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[43,49,73]. Similarly, *Esrp1* and *Esrp2* expression transiently peaked between D6 and D10 for OKMS, and between D14 and D20 for OKS (group 4), following the same pattern as observed with markers of the MET (**Fig. 1D**). This is in line with their known role in promoting AS changes at the MET stage of reprogramming[49,74]. 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 first filtered against lowly expressed isoforms by keeping ones for which expression (TPM $\geq$ 1) is detected in at least 10% of the dataset samples, which culminated to a list of 288,049 isoforms (**Extended Data Fig. 3A** and **Table S3**). We did so to refine our analysis and avoid bias of including splicing events that may arise from lowly represented isoforms. Also, events occurring when comparing NPCs to FBSL samples, which represent changes due to media change only, were removed to focus on events only induced by reprogramming factors (**Extended Data 4**). With this in mind, we detected a total of 44,843 and 85,272 differential isoforms and AS events, respectively, occurring throughout the

reprogramming process (**Extended Data Fig. 3A**). The majority of AS events, with alternative first exons dominating the events, occurred when NPCs transitioned to the early stages of reprogramming (n=9,314 for OKMS; n=8,083 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 or TPM values to obtain differentially expressed isoforms (DEIs), where the number of ΔPSI events or DEIs is proportionally indicative of the number of IS events. Using the ΔPSI events method, we found that IS peaked early during reprogramming (NPCs to early/mid-reprogramming) with equivalent numbers of more and less abundant events (~3,000 in total) (**Fig. 3D** and **Table S3**). 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**). However, events more abundant in early phase reprogramming showed lower coding probability (i.e. decreased or disrupted ORFs), which was more pronounced in OKS reprogramming (**Fig. 3D**). This was consistent with the specific increase in non-coding NNC isoforms in the early phase of OKS reprogramming (**Fig. 2F**). Taken together, this suggests that AS leads to loss of protein-coding potential and function, and tight regulation of non-coding transcripts in the early phase of reprogramming.

5. Selective interplay between gene expression and alternative 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 total, we identified 8,543 genes that undergo splicing changes during reprogramming, and of these 47 genes (among them *Tpm1, Mbnl2, Zfp207, Fgfr1, Palm, Fn1, Dst, Fyn*) were found in common between our dataset and two other reprogramming datasets[49,55] (**Fig. 4A**). However, we found that more than 80% of the genes detected in these datasets overlapped with our dataset, whether the cells used for reprogramming were mouse embryonic fibroblasts (MEFs) (581 genes out of 676, i.e. 86%) or B cells (119 genes out of 148, i.e. 80.4%). This suggested that many AS events are conserved during induction to pluripotency, independently from the starting cell type (**Fig. 4A** and **Table S3**), but also

highlights the strength of using lrRNAseq to detect novel isoforms and splicing events compared to other technologies.

Of the 47 common genes, we found that *Mbnl2*, a negative regulator of AS in pluripotency[43,49,73], 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[75]. Similarly, skipping exon 9 causes a frameshift and decreased protein stability[75] (**Fig. 4B**). Interestingly, exons 6 and 9 were specifically skipped in OKS reprogramming from the middle phase (**Fig. 4C**), suggesting loss of function of MBNL2. This was not observed in OKMS reprogramming, but global *Mbnl2* expression was down-regulated in OKMS mid-reprogramming (**Fig. 4C**, **Fig. 3A** and **Table S3**), suggesting the potential loss of function of MBNL2 is selectively diminished in reprogramming and different mechanisms of gene regulation are selectively used between the two systems (i.e. splicing versus expression changes in OKS and OKMS, respectively). Other commonly spliced genes are involved in cell signaling and adhesion (**Fig. 4D**). Among them are negative regulators of pluripotency, such as tyrosine kinases *Fgfr1* and *Fyn* (**Fig. 4A** and **4D**), which promote differentiation of ESCs[76]. *Fyn*, a SRC family kinase, contains a mutually exclusive exon 7 (exons 7A and 7B, **Fig. 4E**), that leads to differences in *Fyn* SH2 linker and N-terminal of the catalytic domain SH1. Inclusion of exon 7A decreases FYN accessibility to its ligand and thus results in lower activation of the kinase[77]. 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. 4F**). Taken together, these specific examples demonstrate how alternative splicing can selectively regulate the function of genes to ensure that they do not negatively impact reprogramming to pluripotency.

6. Isoform switching selectively regulates non-coding isoforms during early phase reprogramming

Given that both expression and AS selectively regulate reprogramming, we sought to determine if these splicing patterns affected isoform expression. From the 288,098 transcripts identified from our previous isoform classification, we kept only those that are differentially expressed in at least one sample comparison in our dataset. This refined the list to 154,355 transcripts (**Extended Data Fig. 3A** and **Table S4**). Next, we calculated the number of isoforms significantly more or less expressed during the different phases of reprogramming (as determined in **Fig. 1F**) (**Table S4**). Similar to our splicing analysis, we excluded isoforms affected by media change only (FBSL to NPCs comparison, **Extended Data Fig. 5A**). We again found that reprogramming, in both OKMS and OKS, follows a biphasic expression pattern (**Fig. 5A**) as seen with splicing (**Fig. 3C**) and gene-level (**Extended Data Fig. 2B**) analyses, and this pattern was similar whether isoforms were classified as coding or non-coding (**Extended Data Fig. 5B**). Interestingly,

up-regulation of isoforms dominated the early steps of reprogramming (first phase), whereas down-regulation is more prevalent in middle stages, which is a marked difference with gene-level differential expression where equivalent numbers of genes were up- and down-regulated (**Extended Data Fig. 2B**). Both reprogramming systems also showed an up-regulation of isoforms in the late stages (second phase), but OKS reprogramming showed a less pronounced down-regulation of isoforms. This suggests that the early phases of reprogramming are in part conserved in terms of isoform expression regulation. This also coincides with the fact that several molecular and cellular mechanisms of importance for reaching the pluripotent state take place in the early steps of reprogramming[58,62,78–80]. 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") (**Fig. 5B**). This reduced our list of DEIs to 43,529 transcripts (28.2 % of total DEIs) (**Extended Data Fig. 3A** and **Table S4**) but also allowed to define and track isoform expression patterns. For example, we found that the majority of isoforms that were down-regulated in early to mid-reprogramming remained down-regulated throughout reprogramming (n=21,449 for OKMS; n=16,180 for OKS) (**Fig. 5B**).

With this refined DEI list and given that both gene expression and AS are selectively regulated as we previously shown, we now sought to integrate the notion of isoform expression with proportionality in our IS analysis. This is different from our PSI-based IS analysis, 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. 5C** and IS matrix in **Table S4**). We found 1,203 and 1,242 genes undergoing IS for OKMS and OKS, respectively, affecting about 10% of genes where isoforms are differentially expressed between NPCs and mid-reprogramming (**Table S4**). GO-term enrichment analysis showed that genes commonly undergoing IS between NPCs and mid-reprogramming of both OKMS and OKS were associated with processes such as mRNA processing, RNA splicing, translation, cell division, DNA repair, chromatin remodeling, apoptosis, cellular respiration and stem cell population maintenance and differentiation (**Fig. 5D** and **Table S4**). 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 systems. Interestingly, the two main GO-terms for both reprogramming systems were associated with RNA splicing and processing, including genes such as *Srsf2* and *Srsf3*, which are well-known regulators of AS in early reprogramming[48,72]. *Srsf2* gene-level expression did not significantly change during both OKMS and OKS reprogramming (**Extended Data Fig. 5C**). We identified 9 DEIs for *Srsf2*, two coded for the proper protein, six were predicted to undergo NMD, and one non-coding (**Extended Data Fig. 5D**). Of the six NMD transcripts, two were NIC (new

combinations of known splice junctions) isoforms that included a splicing in exon 3 (NMDex3) that was distinct from the protein-coding (main) isoforms (**Extended Data Fig. 5D**). We compared expression profiles and fractions of the main isoforms, the two NMDex3 isoforms and the non-coding isoform. In OKMS reprogramming, the expression and proportion of the protein producing isoforms increased significantly (going from 52.2% in NPCs to 80.6% in late phase, **Fig. 5E**, **Extended Data Fig. 5E** and **Table S4**), while there was a significant decrease in the proportion of NMDex3 isoforms (going from 8.5% in NPCs to 2.2% in middle phase) and the non-coding isoform (going from 13.6% in NPCs to 0% in late phase and subsequent timepoints) (**Fig. 5E** and **Extended Data Fig. 5E**). As for OKS reprogramming, the non-coding fraction decreased significantly (going from 13.6% in NPCs to 0% in subsequent timepoints) but this time the NMDex3 isoforms remained unchanged (around 10%), thus allowing an increase in proportion of the isoforms giving rise to the proper SRSF2 protein (going from 52.2% in NPCs to 86.2% in late phase). All other NMD isoforms also decreased in proportion as reprogramming progressed (going from 25.7% in NPCs to 7.0% in iPSCs) (**Fig. 5E** and **Extended Data Fig. 5E**). Thus, in both reprogramming systems, proper SRSF2 protein isoforms dominate later in reprogramming through expression control of different isoforms, and more specifically by IS of the non-coding isoform to the benefit of the main coding isoforms. This is consistent with changes in coding probability diminishing in early reprogramming phases, to then increase in later phases as observed in our ΔPSI-based IS analysis (**Fig. 3D**).

In the case of *Srsf3*, its gene-level expression remained unchanged in OKMS but decreased during OKS reprogramming (**Extended Data Fig. 5F**). Similar to *Srsf2*, *Srsf3* also undergoes selective isoform expression regulation during reprogramming. We identified 16 isoforms: 3 main isoforms coding for the proper protein, nine had significantly lower coding potential, where most were predicted to undergo NMD (**Extended Data Fig. 5G**), and four NNC (novel splicing sites) isoforms that produced a longer protein product, showed increased coding potential, and lacked the RS domain (**Extended Data Fig. 5G**), which is important for interactions with other proteins and nuclear retention of SRSF3[81]. Again, OKMS and OKS reprogramming showed differences in isoform expression patterns, but in both cases the main isoforms dominated at the end of reprogramming (89.3% in iPSCs). In OKS reprogramming, the proportion and expression of the main isoforms significantly dropped early in reprogramming (going from 82.1% in NPCs to 32.5% in early phase) but increased gradually as reprogramming progressed (**Fig. 5E**, **Extended Data Fig. 5H** and **Table S4**). In contrast, the proportion of isoforms with reduced coding probability increased in the early phase of OKS reprogramming (going from 6.0% in NPCs to 43.6% in early phase) but significantly diminished in the late phases (1.8% in late phase). This opposing pattern of expression to main isoforms suggested loss of function of SRSF3 in early OKS reprogramming that was gained back by switching to the main isoforms in later phases. In terms of OKMS, it was the isoforms with longer CDS (i.e. missing the RS domain) that gradually increased in proportion (going from 12.8% in NPCs to 26.4%

in middle phase) and expression (**Fig. 5E** and **Extended Data Fig. 5H**) and then decreased in later phases (4.6% in late phase), demonstrating differences with OKS regarding which isoforms were regulated early on. However, the outcome of reaching a functional SRSF3 was nonetheless the same between the two reprogramming systems. The regulation of SRSFs suggested again that selective regulation through expression and/or AS allows for production of protein products that lead to proper reprogramming and iPSCs.

Thus far we demonstrated that *Srsf2/3* isoforms with low coding probability or that are non-coding were selectively regulated during reprogramming, and investigation of AS and IS also suggested that coding and non-coding isoforms were regulated differently and at different phases of reprogramming (**Fig. 3D**). Thus, we assessed the respective biotypes of the genes undergoing IS to determine if there was selective pressure for or against non-coding isoforms. We found that around 97% correspond to protein-coding genes (96.84% for OKMS; 97.75% for OKS), whereas only about 1.6% were lncRNA genes (1.58% for OKMS; 1.61% for OKS) and less than 2% constituted pseudogenes or other biotypes (1.58% for OKMS; 0.64% for OKS) (**Fig. 5F**). Strikingly, when comparing these numbers to the proportion of biotypes in the list of genes undergoing differential isoform expression but not necessarily IS (n=9,788 genes for OKMS; n=9,835 genes for OKS), we found that lncRNAs were significantly less enriched within the IS gene list: Indeed, lncRNAs were expected to amount for 6-7% of genes (6.1% for OKMS; 7.2% for OKS), but they represented only 1.6% in our OKMS and OKS datasets for IS (**Fig. 5F**). In contrast, protein-coding genes were significantly more enriched within the IS gene list (96.8% for OKMS; 97.8% for OKS) compared to expected values (91.7% for OKMS; 90.9% for OKS). These results suggest that there is a negative selection against lncRNAs to undergo IS. With this in mind, we postulated that lncRNAs that undergo IS may harbor important functions during the first half of the reprogramming process and thus for pluripotency acquisition. This was also consistent with our results demonstrating a selective pressure and tight regulation of non-coding transcripts during the early phases of reprogramming (**Data Fig. 2F**).

7. Identification of a candidate non-coding transcript regulating the pluripotent state

To reduce false-positive discovery of genes, more specifically lncRNA genes, that undergo IS and to increase chances of identifying functionally relevant IS events, we merged our list of ΔPSI isoforms (n=7,379 genes) and DEIs (n=2,021 genes) comparing NPCs to mid-reprogramming in either OKMS or OKS systems (**Fig. 6A**). We obtained 972 candidates that were common to these two lists (**Fig. 6A** and **Table S5**). Among them, we identified 19 lncRNA candidates (**Fig. 6A** and **Table S5**). In order to validate the importance of these candidates for reprogramming, we cross-referenced our list with lncRNAs that were reproducibly and differentially expressed in other reprogramming systems[23,38,82] (**Fig. 6B**). Of the 19 lncRNAs, 9 were identified and expressed in these other datasets (**Fig. 6B** and **Extended Data Fig. 6A-B**).

From these, we focused on lncRNAs that were highly expressed at the iPSC stage because we aimed at identifying new genes potentially involved in the regulation of pluripotency. From this, we obtained two annotated candidates: *2700038G22Rik* and *2010204K13Rik* lncRNAs (**Fig. 6B**), and a human syntenic ortholog could be found for both. However, the human ortholog for *2010204K13Rik* lncRNA appeared to be the coding gene *PAGE4*, which was only expressed in late reprogramming (**Extended Data Fig. 6C**). *2700038G22Rik* lncRNA was the only candidate meeting all the above criteria without ambiguity, justifying the focus on this lncRNA for further analysis. Expression of *2700038G22Rik* strongly increased during reprogramming towards iPSCs (**Fig. 6C**), and this was reproducible in different datasets from 2º MEFs and 2º tail-tip fibroblasts (TTFs), with expression levels in iPSCs comparable to those in control ESCs[23,28,82] (**Fig. 6D**). Moreover, expression of the human ortholog of *2700038G22Rik*, *SNHG26*, showed a similar expression pattern of up-regulation during human reprogramming[83] (**Fig. 6E**). This suggests that our results were conserved in reprogramming in general.

Because *2700038G22Rik* (named *Snhg26* for here on) was identified as undergoing IS, we analyzed its different isoforms and identified 11 transcripts (**Fig. 6F**). Only four could be identified in the latest Ensembl database version (release version 113.39, Aug. 2024)[84] (**Fig. 6F**), highlighting dataset specific isoform diversity and the importance of long-read data integration into public databases. In our reprogramming dataset, three isoforms showed both differential PSI and differential isoform expression during reprogramming (**Fig. 6G** and **Extended Data Fig. 6D-E**). However, 2 isoforms, *G38722.217* (*Ensmust00000181574*) and *G38722.216* (*Ensmust00000268635*), demonstrated an opposite pattern of expression and clear isoform switching between NPC stage and mid-reprogramming of both OKMS and OKS systems (**Fig. 6G**). Expression of *G38722.217* increased drastically (relative to NPCs for OKMS and to FBSL for OKS reprogramming), while expression of *G38722.216* was down-regulated. The conserved pattern of expression among different reprogramming systems and species suggests that *Snhg26*, more specifically the *G38722.217* isoform, is an important regulator of the early stages of pluripotency induction.

8. *SNHG26* lncRNA enhances the reprogramming route to pluripotency

In order to investigate if *Snhg26*, isoform *G38722.217*, could have a function during the acquisition of pluripotency, we performed mouse reprogramming experiments where we overexpressed either a control RNA or *Snhg26* with the reprogramming factors OKS. Upon DOX addition, the proportion of reprogramming mCherry+ cells peaked earlier for OKS-*Snhg26* (80% of cells at D8) than for OKS-control (58.35% of cells at D14) (**Fig. 7A**, left and **Extended Data Fig. 7A-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*) and increasing similarly between the conditions to reach their maximum at D20 (21.78% for OKS-control; 18.50% for OKS-*Snhg26*, **Fig. 7B**, left). By this time, most cells had spontaneously silenced transgenes

expression (11.05% of mCherry+ cells for OKS-control; 13.39% for OKS-*Snhg26*) (**Fig. 7A**, left). Surprisingly, the increased number of mCherry+ cells when *Snhg26* was expressed did not lead to significantly more GFP+ cells upon DOX removal (63.35% for OKS-control; 75.55% for OKS-*Snhg26* at D14, **Fig. 7B**, right), even if mCherry levels remained high (going from 51.61% at D6 to 73.4% at D10 for OKS-*Snhg26,* compared to 37.70% at D10 for OKS-control, **Fig. 7A,** right). 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*, **Fig. 7A,** right). The fact that the proportion of mCherry+ cells for OKS-*Snhg26* 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* expression increased the number of reprogramming cells: 1) *Snhg26* could directly affect proliferation of cells that commit to reprogramming but not 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* may cause increases in iPSC colony formation (i.e. more cells commit to iPSCs). For the latter scenario, the total number of cells would be greater under *Snhg26* expression because of increase in reprogramming cells (i.e. mCherry+ cells); thus, looking at percentages will not reflect this increase as 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 Data 2** for details).

To test these two scenarios, we first addressed whether *Snhg26* expression could increase the proliferation of reprogramming cells. We found that *Snhg26* expression significantly and consistently increased the population of mCherry+ cells (297,253 cells for OKS-control; 471,693 cells for OKS-*Snhg26* at 120 hours post-DOX induction) (**Fig. 7C**, left) and the total number of cells (**Extended Data Fig. 7C**). However, no significant difference was observed in mean doubling time of these mCherry+ or total populations between 48h and 120h of DOX induction between control and *Snhg26* conditions (**Fig. 7C**, right, and **Extended Data Fig. 7C**), suggesting that *Snhg26* 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* RNA, in NPC medium and independent of reprogramming medium conditions. Once again, their average proliferation rates were similar (**Extended Data Fig. 7C**), suggesting that *Snhg26* lncRNA overexpression did not impact the growth of the starting NPC lines. Finally, to make sure that *Snhg26* expression does not influence the division rate of pluripotent cells themselves, we compared the doubling times of ESCs overexpressing *Snhg26* 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 wild-type ESCs to *Snhg26*

lncRNA overexpression (**Extended Data Fig. 7C**). These results demonstrated that *Snhg26* 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 lead 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* effect on the process (**Supplementary Data 2**). 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* (3,823,884 cells/mL) conditions (**Fig. 7D**). It subsequently decreased, rather quickly, to give rise to GFP+ cells, but interestingly, there were more iPSCs cells at D20 in the OKS-control condition (2,195,702 cells/mL) than when *Snhg26* lncRNA was overexpressed (1,391,328 cells/mL) (**Fig. 7E**), most likely because most cells were not able to silence the transgenes in OKS-*Snhg26* reprogramming. However, 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*) (**Fig. 7E**, right). This was accompanied by an increase in mCherry/GFP negative cells (**Extended Data Fig. 8A**) highlighting that scenario 2 is the most plausible explanation for our observations and that *Snhg26* 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* 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* was expressed compared to the control condition regardless of reprogramming factors order, cell origin, or whether it was done using transient or stable transfections (**Extended Data Fig. 8B-G**). Together, these results validated that *Snhg26* significantly enhanced the induction of iPSCs in various reprogramming systems.

We next asked if *Snhg26* could influence the frequency of cells able to engage into reprogramming by changing cell identity and enhancing the cells' intrinsic rate of reprogramming. To assess this, we calculated the frequency of cells engaging in reprogramming and reaching the iPSC stage (GFP+) by performing 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 Data 3** for details)[85–87]. 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 (**Fig. 7F**). 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). However, when we

overexpressed *Snhg26* lncRNA 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. 7G**). Together, these results led us to conclude that *Snhg26* did not accelerate the reprogramming kinetics via faster cell division but rather enhanced the overall efficiency by increasing cell-intrinsic mechanisms (for a more detailed explanation, see **Supplementary Data 2**).

To validate that iPSCs generated with or without *Snhg26* overexpression were *bona fide* iPSCs, we isolated clones from the OKS-control and OKS-*Snhg26* reprogramming experiments and found them to be able to contribute to the formation of chimeras when aggregated with mouse wild-type tdTomato 8-cell morula embryos. Indeed, ten out of 15 E8-8.5 embryos showed high or partial contribution for OKS-control, as visualized with the negative cells/area in tdTomato-positive embryos (**Extended Data Fig. 8H**). As expected, iPSC contribution was also limited to the embryonic area (i.e. epiblast derivatives), with extraembryonic tissues remaining tdTomato-positive. This confirmed that our OKS reprogramming system gives rise to *bona fide* iPSCs. For OKS-*Snhg26* condition, two embryos out of five E8-8.5 embryos showed high contribution (**Extended Data Fig. 8H**), confirming that *Snhg26* expression during reprogramming gave rise to *bona fide* iPSCs.

9. *SNHG26* lncRNA is a developmental gene and is necessary for maintenance of both mouse and human pluripotent state of ESCs

To investigate further the regulation and function of *Snhg26* in pluripotency, we used our previous ChIPseq during reprogramming of 2˚MEFs to iPSCs[23] and analyzed the chromatin landscape around *Snhg26* locus. The *Snhg26* promoter was marked by H3K4me3 and H3K27me3 and devoid of DNA methylation in 2˚MEFs but lost the H3K27me3 mark as cells progressed through reprogramming towards iPSCs (**Fig. 8A**), suggesting a bivalent locus that is activated upon loss of the repressive mark H3K27me3. This is in accordance with up-regulation of *Snhg26* during reprogramming and a sign of a developmentally regulated genes[88,89]. Moreover, a previous study identified *Snhg26* as part of a cohort of ESC-specific lncRNAs and showed that knock-downs of KLF2 or N-MYC pluripotency-associated transcription factors, both lead to a down-regulation of *Snhg26* (named *linc1533* in that study) in mouse ESCs[38]. They also showed the binding of KLF4, C-MYC and N-MYC to *Snhg26* promoter[38], suggesting that *Snhg26* may be regulated by the pluripotency-transcriptional network.

Because those results suggested a role for *Snhg26* in ESCs, and because *Snhg26* was selected from a lncRNA candidate list for its conserved expression pattern during mouse and human reprogramming, we wondered if *Snhg26* expression could be conserved in the broader context of pluripotency and development. To this end, we compared expression of total *Snhg26* lncRNA and its human counterpart, *SNHG26* lncRNA, during mouse[16] and human[90] embryo development using previously published datasets (**Fig. 8B**).

Interestingly, *Snhg26* expression peaked at the 4-cell and 8-cell stages of the mouse and human pre-implantation embryos, respectively, and remained expressed up to the pluripotent state, albeit at a lower expression level. This peak expression corresponds to the zygotic genome activation stages in the respective species and suggests a role for *Snhg26* in embryo development (**Fig. 8B**). Furthermore, if we examine human and mouse *Snhg26* expression during *in vitro* differentiation of ESCs towards cortical neurons[91,92], we observe that their expression drops drastically from the first days of the differentiation process (**Extended Data Fig. 9A**). These results suggest that expression of *Snhg26* and its human counterpart is specific to early embryo development and the associated pluripotent state. Finally, because mouse and human ESCs correspond to different states of pluripotency[93,94], we examined *SNHG26* expression in primed and *naïve* human ESCs (hESCs), as well as *in vitro* equivalent of eight cell-like cells[95]. We found that *SNHG26* was highly expressed in *naïve* and eight cell-like cells compared to primed hESCs, which is in keeping with results from pre-implantation embryo datasets (**Fig. 8C**). Thus, the genomic conservation, identical expression profiles over several biological systems, and bivalency and regulation of its locus by pluripotency-associated transcriptional factors suggested that it could share a conserved functional role in pluripotency regulation.

To address its role in pluripotency, we first sought to determine where *Snhg26* lncRNA is localized as localization of a lncRNA is generally indicative of its function[96,97]. We found *Snhg26* to be nuclear (83.8% on average) according to quantifications performed after cellular fractionation of mouse ESCs (**Extended Data Fig. 9B** and **Table S5**). 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* transcript demonstrated strong co-localization with DAPI staining (**Extended Data Fig. 9C**), further validating *Snhg26* nuclear localization in mouse ESCs. Therefore, we knocked-down *Snhg26*'s transcription in mouse and human ESCs by targeting its promoter with a CRISPR-interference (CRISPRi) system[98,99], which ensures that all its transcripts were targeted. Indeed, *SNHG26* has 54 human isoforms according to the latest version of Ensembl (release 113.39, Aug. 2024)[84], with only 10 being supported by minimal evidence (**Extended Data Fig. 9D** and **Table S5**). Thus, using the CRISPRi system is justified as CRISPRi enables us to knock-down all the transcripts potentially expressed in ESCs to assess the overall function of a lncRNA. Knock-down of *Snhg26*'s transcription for 5 days in mouse ESCs led to down-regulation of pluripotency genes *Oct4, Nanog* and *Tfcp2l1* (**Fig. 8D**). Contrary to results in mouse ESCs, in hESCs, no significant decrease was observed for the core pluripotency genes *OCT4* and *NANOG* (**Fig. 8E**). This can be explained by the fact that, as previously mentioned, mouse and human ESCs do not correspond to the same state of pluripotency. 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) (**Fig. 8E**). These results suggested that similarly to its mouse ortholog, *SNHG26* was necessary to maintain the pluripotent state of hESCs. Altogether, the effects of overexpression of *Snhg26* on reprogramming efficiency and of its knock-down in ESCs demonstrate its role in the regulation of the pluripotent state. Thus, lrRNAseq analysis of our optimized reprogramming systems enabled us to identify a new lncRNA whose function is conserved in regulating the maintenance of the pluripotent state (**Fig. 8F**).

Discussion

Previous studies have demonstrated that AS regulation is an important factor in somatic cell reprogramming to pluripotency. Two reports[49,55], 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 are *Srsf2* and *Srsf3* (**Extended Data Fig. 5**), which were also identified by these previous studies and others[45,48] as important AS regulators during reprogramming and pluripotency. Expression of non-coding isoforms, or isoforms with low coding probability, 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 exposed 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 that recapitulates known features of reprogramming in terms of timelines, gene expression profile and alternative splicing events, but with an improved efficiency. Running lrRNAseq at different timepoints of reprogramming, we brought to light the underestimated diversity of isoforms expressed during this process, with about 80% of the total detected isoforms being previously unannotated and many originating from new splice junctions (i.e. “Novel-Not-in-Catalog” transcripts) (**Fig. 2A**). This reinforces the utility of lrRNAseq to address the shortcomings of

small reads RNAseq in identifying full-length isoforms. Closer examination of NNC isoforms showed that the transcripts that were classified as non-coding were regulated in the early and middles 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 spliced isoforms had a lower coding potential before this switch (**Fig. 3D**), 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, together with an observed enrichment of isoforms from new genic and intergenic isoforms at the transition to the late and iPSC phases of reprogramming (**Fig. 2F**), 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 [38,100–104], 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 *Snhg26* lncRNA. *Snhg26* overexpression greatly increased the number of reprogramming cells and *bona fide* iPSCs (**Fig. 7**). Furthermore, knock-down of *Snhg26* in mouse and human ESCs down-regulated expression of several pluripotency markers (**Fig. 8D-E**). This validated the conserved role for this new transcript in the regulation of pluripotency. Moreover, the conserved expression pattern of mouse and human *Snhg26* during embryo development (**Fig. 8B**) and their common role in maintaining the pluripotent state is of interest, as it illustrates other biological contexts in which the *Snhg26* lncRNA is evolutionary conserved. Indeed, a recent study[105] 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.

In summary, it is important to note that only by analyzing isoform switching of *Snhg26* that we were able to identify the 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 pluripotency.

Limitations of the study

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. Also, we only investigated *Snhg26*'s effect on reprogramming from NPCs, when standard reprogramming models usually use MEFs. However, the expression profile of *Snhg26* from MEFs and TTFs found in other datasets support the idea that it could enhance reprogramming from other cell types as well. We did not investigate whether *SNHG26* could enhance human reprogramming, so we cannot state that its effect on pluripotency induction is conserved. However, given that its knock-down affects human ESCs we suspect that loss of *SNHG26* will negatively impact human iPSC reprogramming.

Authorship

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

Acknowledgments

We thank Dr Ian Roger for advice and reading of the manuscript. Gaëlle Bourriquen assisted with the mouse CRISPRi sgRNA design. This work was funded by the Canadian Institutes of Health Research (CIHR) to S.H (PJT-378019) and supported by the Fonds de recherche du Québec (FRQ) through the Research Centre Grant. V.F. is a recipient of a training award of the Fonds de recherche du Québec - Santé (FRQS). G.K. is a recipient of training awards form the Natural Sciences and Engineering Research Council of Canada (NSERC) and the FRQS. V.W. is a recipient of a training award of the Fonds de recherche du Québec - Nature et Technologies (FRQNT). S.H. is a Junior 2 Research Scholar of the FRQS.

Competing interests

The authors declare no competing interests.

Materials and methods

Generation of the plasmid constructs

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 a PB-TAC-OKMS to replace the OKMS sequence) and PB-TAC-OSKM (Addgene #80481) containing a doxycycline-inducible TetO promoter (composed of a Tetracycline Response Element followed by a minimal CMV 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[106] 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 (GOI) 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, i.e. a fusion between a reverse tetracycline repressor and a VP16 transcription factor) and a puromycin (P) or neomycin (N) resistance gene. GOI were inserted into PB-TetO-AIO destination vectors by Gateway™ cloning (ThermoFisher Scientific #11789020 & #11791020) strategy according to the manufacturer's instructions.

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 #11789020 & #11791020) strategy according to the manufacturer's instructions. Single guide RNAs (sgRNAs) were designed using the CRISPR 10K Tracks of the UCSC genome browser[107] and ordered from IDT™. They were resuspended and annealed before their ligation into an in-house PB-U6-Gateaway plasmid beforehand digested with BbsI-HF® (NEB #R3539) using a T4 DNA ligase (ThermoFisher Scientific #EL0016) according to the manufacturer's instructions (sgRNA sequences are listed in **Table S5**). The sequences of all entry clone inserts 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* long non-coding RNA (lncRNA) & 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 201 (*ENSMUST00000181574*) were used with the Deoxynucleotide (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.

Similarly, 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 entry clone inserts 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 **Table S5**).

Mouse ESC culture

ESC lines, including R1 (RRID: CVCL_2167) and Oct4-GFP (R1 ESCs harboring a *GFP* reporter gene under the control of the *Oct4* gene promoter from Viswanathan *et al.*[108]) (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 Leukemia Inhibitory Factor (LIF, home-made from pGEX-2T-MLIF, [109]) at 20ng/mL, 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 washed with DPBS (ThermoFisher Scientific #14190250) and treated with TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) for 5 minutes. Then, TrypLE™ Express Enzyme was neutralized using DMEM (high glucose, ThermoFisher Scientific #11965118) with 5% KnockOut™ Serum Replacement (KSR, ThermoFisher Scientific #10828-028) and cells were 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 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) and 10^{-4}M 2-mercaptoethanol (ThermoFisher Scientific #21985023). The resulting primed cells were passaged at day 2 (D2) and day 4 in FBS-LIF 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 FBS-LIF medium on 0.1% porcine gelatin. Twenty-four hours later, medium was switched with 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 washed with DPBS (ThermoFisher Scientific #14190250) and treated with StemPro™ Accutase™ Cell Dissociation Reagent (ThermoFisher Scientific #A1110501) for 5 minutes. Then, StemPro™ Accutase™ Cell Dissociation Reagent was neutralized using DMEM (high glucose, ThermoFisher Scientific #11965118) with 5% KSR (ThermoFisher Scientific #10828-028) and cells were 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 washed with DPBS (ThermoFisher Scientific #14190250) and treated with Dispase (StemCell Technologies #07923) for 5 minutes. Then, Dispase was washed away using DMEM/F-12 (ThermoFisher Scientific #11330057) and cells were 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.

Cell transfection in mouse ESCs and NPCs

ESC lines cultured in a humidified incubator at 37°C under 5% CO2 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 cultured in a humidified incubator at 37°C under 5% CO2 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. Cells were then selected with 1µg/mL puromycin (Bio Basic #PJ593) for 4 days or 400µg/mL G-418 sulfate (Wisent #400-130-IG) for 6 days.

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.

For CRISPRi cell lines, 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.

Cell transfection in human ESCs

Primed hESC lines cultured in a humidified incubator at 37°C under 5% CO2 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.

Reprogramming towards induced pluripotent stem cells (iPSCs)

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 FBS-LIF 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-LIF or 2i-LIF medium and passaged every 2-4 days depending on colonies density. For passaging, cells were washed with DPBS (ThermoFisher Scientific #14190250) and treated with TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) for 5 minutes. Then, TrypLE™ Express Enzyme was neutralized using DMEM (high glucose, ThermoFisher Scientific #11965118) with 5% KSR (ThermoFisher Scientific #10828-028) and cells were 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 iPSC pluripotency in iPSC-embryo chimeras

An iPSC clamp consisting of 4-5 cells of a clone from either the OKS + control or OKS + *Snhg26* lncRNA cell line were aggregated with two WT 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 in an incubator 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.

Doubling time assays

ESC or NPC cell lines were seeded in replicates at 5 x 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 FBS-LIF 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 by one-tailed t-test (two samples with unequal variance). n.s. ≥ 0.05, * < 0.05.

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 FBS-LIF 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 by one-tailed t-test (two samples with unequal variance). * < 0.05.

Flow cytometry

Cells were harvested using TrypLE™ Express Enzyme (ThermoFisher Scientific #12604021) for dissociation, washed with DPBS (ThermoFisher Scientific #14190250), 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: $density\ of\ mCherry\ cells = \%\ of\ live\ single\ mCherry\ cells \times (\frac{number\ of\ live\ single\ cells}{cytometer\ uptake\ volume})$. 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 by one-tailed t-test (paired samples). n.s. ≥ 0.05, * < 0.05.

ESC nuclear/cytoplasmic fractionation

Fractionation was performed following Gagnon *et al.*[110]. Briefly, R1 ESCs cultured in a humidified incubator at 37°C under 5% CO2 and 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) and washed with ice-cold DPBS (ThermoFisher Scientific #14190250) before counting. 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. Samples were then processed by RT-qPCR. *Snhg26* lncRNA localization was evaluated by RT-qPCR by calculating the % 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 **Table S5**). The average values of three independent experiments were used and represented with the STD using GraphPad Prism (v7.04). P-values were calculated by two-tailed t-test (two samples with unequal variance).

RNAScope *in situ* hybridization

R1 mouse 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). For all pictures, 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* and *SNHG26* lncRNAs

The inducible knock-down cell lines were generated using CRISPRi-based methods following Gilbert *et al.*[99]. 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. 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 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 **Table S5**). The average values of 2-3 independent experiments were used and represented with the STD using GraphPad Prism (v7.04). P-values were calculated by one-tailed t-test (two samples with unequal variance). * < 0.05, ** < 0.01, **** < 0.0001.

ESCs 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).

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 2ΔΔCt 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 **Table S5**).

Library preparation and long-read RNA-sequencing (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 **Table S1**). 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), 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 two 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 using 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 reference genome with Minimap2[111] (v2.24-r1122) using the following parameters: -aLx splice –cs = long. Aligned reads were quality-controlled with the NanoPlot software[112] (v1.43.2).

LrRNAseq data analysis of gene expression

Raw read counts were obtained with the featureCounts tool from the Subread package (v2.0.6), using the exon counting mode[113]. Together, the two sequencing runs (35 samples in total; **Table S1**) generated 64,745,442 quality control-passed and filtered reads in total, with an average of 1,849,870 reads and 3.5 Gb of data per replicate (**Table S1**). EdgeR R-package[114] (v3.42.4) was then used to normalize the data, calculate RNA abundance at the gene level (as counts per million reads (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 used for statistics, as described in EdgeR user guide. Differential gene expression was established as genes with a fold change ≥ 1.5 (up-regulation) or ≤ 0.67 (down-regulation) and FDR ≤ 0.05. Gene expression plots were generated using Microsoft Excel by averaging the log10(CPM+1 expression values) of different gene lists and representing the SEM as a shaded contour (gene lists are listed in **Table S1 & Table S3**). GO term enrichment was assessed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (v2022q3)[115]. Biological Processes (BP) annotations were analyzed with a cut-off of FDR ≤ 0.1 for significant enrichment. Hierarchical clustering analysis of whole transcriptome was performed using the Ward.D2 method in the stats R package[16] (v4.3.1) to separate RNAseq samples into reprogramming groups. 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.

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)[117], 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 SQANTI 3 software[66] (v5.1.1) using the sqanti3_qc.py script and the “--CAGE_peak” and “--polyA_motif_list” options to identify TSS and TES, respectively. Briefly, the sqanti3_qc.py script (with the “--CAGE_peak” and “--polyA_motif_list” options to identify TSS (transcriptional start site) and TES (transcriptional end site), respectively) was used to classify isoforms using GRCm38/mm10 as reference genome and Ensembl’s gene annotations version 99 as reference transcriptome. Isoform 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 TSS and polyA 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 (see Extended Data Fig. 3A). Percentages of total, coding and non-coding isoforms in each structural categories and subcategories were calculated based on this total count of 1,038,414 unique isoforms. The coding potential score of detected transcripts was assessed with the Coding-Potential Assessment Tool (CPAT)[118] (v3.0.5), which accounts 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 one-tailed t-test (two samples with unequal variance). n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

Analysis of isoform level expression and isoform switching

Isoform level expression was quantified using IsoQuant[119] (v3.3.0) using default parameters, and low expression isoforms (isoforms where expression is ≥ 1 tpm in at least 3 samples (i.e. 10% of samples from the whole dataset)) were further filtered to obtain a total of 288,098 isoforms. EdgeR R-package[114] (v3.12.1) was then used to normalize the data, calculate differential transcript expression, and perform statistical analysis. Isoforms with a p-value ≤ 0.05 in at least one reprogramming phases comparison were kept. This represented 154,355 isoforms that were selected for further analysis (see Extended Data Fig. 3A). Differentially expressed isoforms (DEIs) were established as transcripts with a fold change ≥ 1.5 (up-regulated) or ≤ 0.67 (down- regulated) and p-value ≤ 0.05 between reprogramming phases. 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. We obtained 43,529 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 during reprogramming. 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 (**Table S4**). Observed percentages of genes undergoing IS among each biotypes were calculated based on the total count of genes undergoing IS between D0 and the middle phase of reprogramming. Expected percentages were them based on the total count of genes with DEIs during reprogramming, and P-values were calculated by Chi-squared test. **** < 0.0001. Gene Ontology enrichment analysis of genes associated with at least one significant IS event per comparison was performed using the Database for Annotation, Visualization and Integrated Discovery ((DAVID) v2022q3)[115]. BP annotations were analyzed with a cut-off of FDR ≤ 0.1 for significant enrichment.

Analysis of alternative splicing

The occurrence of local alternative splicing (AS) and transcript events during reprogramming was assessed using SUPPA2[120] (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 288,098 isoforms was used as input for these analyses. To calculate the percent spliced in (PSI) values for each isoforms 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 transcripts per million reads (TPM) values from Isoqaunt 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 381,236 AS events to 85,272. Differential splicing between reprogramming phases was then established as isoforms with a ΔPSI ≥ 0.1 (more inclusion of an AS event) or ΔPSI ≤ -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 ΔPSI ≥ 0.1 (more abundant isoform) or ΔPSI ≤ -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 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 detected in this and other reprogramming datasets[49,55] was presented as a Venn diagram using InteractiVenn[121].

Visualization of gene loci and alternatively spliced isoforms

Validations of gene loci, alternative splicing events, isoform abundance and isoform switch were performed using the Integrative Genomics Viewer (IGV)[122] (v2.16.2) tracks.

Identification of candidate lncRNAs

Final lists of genes for which alternatively spliced isoforms were significantly more or less relatively abundant (see "Analysis of alternative splicing" section), and genes that undergo isoform switching (see "Analysis of isoform switching" section), between D0 and the middle phase of reprogramming were generated. Overlap between these two lists was presented as a Venn diagram using Microsoft Excel.

Expression of candidate lncRNAs from publicly available datasets were represented with SEM using Microsoft Excel or GraphPad Prism (v7.04). P-values between the frequencies were calculated by two-tailed t-test (two samples with unequal variance). n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

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 **Table S5**). Long-read RNA sequencing raw data, filtered aligned reads and processed data, including transcriptome annotation and expression level matrix, from this study are publicly available and have been deposited in the GEO database under the accession code: GSE282319.

Biological materials availability

All cell lines and biological materials generated in this study are available from the corresponding author upon direct request.

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.

Figure legends

Figure 1: Designing an efficient reprogramming system

A) Schematic of the experimental reprogramming pipeline: NPCs harboring a *GFP* reporter under the control of the *Oct4* gene promoter (green) were transfected with OKMS or OKS reprogramming factors (orange) driving the expression of an *mCherry* reporter (pink). 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 (blue) 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 GFP+ iPSCs and subsequent silencing of the transgenes. Samples were collected at different timepoints (black dots) 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)); FBSL: FBS-LIF medium; 2iL: 2 inhibitors + LIF medium.

B) Plots of logged expression as counts per million (CPM) for immediate responder, intermediate reprogramming, early ESC-like and ESC-like genes (gene lists from Hussein *et al.*[23]) or mesenchymal-to-epithelial transition (MET) genes in OKMS (green) or OKS (blue) reprogramming. Lines represent average expression per timepoint and shaded areas represent SEM. ESC: Embryonic Stem Cell. Also see Extended Data Fig. 2A.

C) Hierarchical clustering on whole lrRNAseq transcriptome of OKMS (green) and OKS (blue) reprogramming samples (clustering method: Ward.D2). rep.: replicate; repro.: reprogramming; PSC: Pluripotent Stem Cell. Also see Extended Data Fig. 2C.

Figure 2: lrRNAseq reveals unprecedented isoform diversity and that non-coding isoforms are exclusively regulated during reprogramming

A) Pie chart and bar graph representing 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: Not-In-Catalog; NNC: Novel-Not-in-Catalog; AS: Antisense.

B) Schematic of isoform structural categories and subcategories, according to SQANTI3 classification[66]. TSS: Transcription Start Site; Comb.: Combination.

C) Bar graphs representing the proportion of isoforms in each structural subcategories (left) or the proportion of coding isoforms going through (blue) or avoiding (grey) nonsense-mediated decay (NMD) among the different structural subcategories of the filtered dataset. Alt.: Alternative.

D) Box plots showing the coding probability of isoforms among the different structural categories of the filtered dataset. 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[118].

E) Bar graph of mean percentages of novel isoforms from novel genes detected in NPCs (n=8), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) of the filtered dataset. Error bars represent the standard deviation (STD) and each dot represents replicates. P-values were calculated by two-tailed t-test with unequal variance. * < 0.05, ** < 0.01. Also see Extended Data Fig. 3D.

F) Heat map (left) showing significant variations in percentage of isoforms across structural categories throughout OKMS and OKS reprogramming (increases: pink, decreases: blue), for either coding or non-coding isoforms. Bar graphs (right) of mean percentages of coding or non-coding NNC isoforms detected in NPCs (n=8), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) of the filtered dataset. Error bars represent the STD and each dot represents replicates. P-values were calculated by two-tailed t-test with unequal variance. n.s. ≥ 0.05, ** < 0.01, *** < 0.001.

Figure 3: Alternative splicing follows a biphasic pattern during reprogramming

A) Heat map (left) and plots (right) depicting logged expression in CPM of pluripotency-related regulators of alternative splicing (n=19) (gene lists from Cieply *et al.*[49] and Vivori *et al.*[55]) in OKMS (green) or OKS (blue) 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 SUPPA 2 classification[120].

C) Splicing events-based quantification of alternative splicing significantly more (blue) or less (turquoise) included at timepoint transitions in OKMS (top) or OKS (bottom) 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; MX: Mutually exclusive exons; RI: Retained Intron; A5': Alternative 5' splice site; A3': Alternative 3' splice site. Also see Extended Data Fig. 4A.

D) Isoform-based quantification (left) of alternative splicing significantly more (blue) or less (turquoise) abundant at timepoint transitions in OKMS (top) or OKS (bottom) reprogramming. Significance based on ΔPSI ≥ 0.1 (more abundant) or ΔPSI ≤ 0.1 (less abundant) and an adjusted p-value ≤ 0.05. Horizontal dotted line represents the inflexion point in the biphasic alternative splicing pattern observed in reprogramming. Also see Extended Data Fig. 4B. Box plots (middle and right) of 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[118]. Black bars represent the median and colored bars represent the averages. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

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

A) Venn diagram representing the overlap between the number of genes undergoing significant alternative splicing detected between our lrRNAseq dataset (reprogramming from NPCs, n=8,543) and the datasets from Cieply *et al.*[49] (reprogramming from MEFs, n=676) and Vivori *et al.*[55] (reprogramming from B cells, n=148). MEF: Mouse Embryonic Fibroblast.

B) Schematic highlighting two alternative splicing events of exons 6 (pink) and 9 (green) within the *Mbnl2* gene locus. Exon skipping of exon 6 (left) or 9 (right) leads to mis-localization or degradation of the MBNL2 protein, respectively. bp: base pair; 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) Plots (left) depicting the PSI values of *Mbnl2* exon 6 (middle) or exon 9 (right) inclusion in phases of OKMS (green) or OKS (blue) reprogramming. Lines represent average expression and shaded areas represent SEM. Empty circles distinguish the FBSL data from the early phase data. Bar graph (right) representing mean expression in transcript per million (TPM) of *Mbnl2* gene in NPCs (n=8), FBSL samples

(n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

D) Bar graph of logged p-values for GO-term enrichment analysis (biological processes and KEGG pathways) for the genes (n=47) common from our lrRNAseq dataset to the datasets from Cieply *et al.*[49] and Vivori *et al.*[55] . FDR ≤ 0.1.

E) Schematic highlighting the mutually exclusive alternative splicing of exons 7a and 7b (pink) within the *Fyn* gene locus. Inclusion of exon 7a leads to a decrease in FYN's ligand accessibility and thus to less FYN kinase activation. SH: Src Homology domain; U: Unique domain.

F) Bar graph (left) representing mean expression in TPM of *Fyn* gene in NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM. Plot depicting the PSI value of *Fyn* exon 7a (right) inclusion in phases of OKMS (green) or OKS (blue) reprogramming. Lines represent average expression and shaded areas represent SEM. Empty circles distinguish the FBSL data from the early phase data. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001.

Figure 5: Isoform switching selects against lncRNAs during early phase reprogramming

A) Bar graphs of total number of isoforms significantly up- (red) or down- (blue) regulated between the different phases of OKMS (left) and OKS (right) reprogramming among the total isoforms (n=154,355) detected in the filtered dataset. Significance based on log2(fold change) ≥ 2 for up-regulated isoforms or ≤0.5 for down-regulated isoforms with p-values ≤0.05. DEIs: Differentially Expressed Isoforms. Also see Extended Data Fig. 5A-B.

B) Plots depicting average expression in TPM for isoforms being significantly differentially expressed between D0 and the middle phase of reprogramming, in NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM.

C) Schematic defining the difference between differential transcript usage (i.e. isoform switching) and differential transcript or gene expression. DTU/IS: Differential Transcript Usage/Isoform switch; DTE: Differential Transcript Expression.

D) Bar graphs of logged p-values from GO-terms enrichment analysis (biological processes) for genes undergoing isoform switch between D0 and the middle phase of OKMS (n=1,203 genes, top) or OKS (n=1,242 genes, bottom) reprogramming. FDR ≤ 0.1. Neg./Pos. reg.: Negative/Positive regulation; pop.: population.

E) Pie charts representing the proportion of each *Srsf2* (top) or *Srsf3* (bottom) transcript categories in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01, *** < 0.001. 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. CDS: Coding Sequence; CP: Coding Potential. Also see Extended Data Fig. 5C, 5E, 5F and 5H.

F) Pie charts representing the proportions of each gene biotype among the genes undergoing isoform switching, according to GENCODE annotation[123]. 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. PCG: Protein-Coding Gene.

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

A) Venn diagram showing overlap between the number of genes and lncRNA genes (in parenthesis) which undergo isoform switch between D0 and the middle phase of our reprogramming systems, based on either the PSI method (n=7,379 genes; n=737 lncRNAs) or the expression method on the other side (n=2,021 genes; n=32 lncRNAs).

B) Schematic showing the pipeline used to isolate lncRNA candidate genes potentially involved in the regulation of pluripotency induction: 19 lncRNA candidates were crossed with lncRNAs identified in a RNAseq dataset from 2° mouse embryonic fibroblasts reprogramming[23], to obtain 9 candidates.

C) Bar graph representing the mean expression in TPM of *2700038G22Rik* gene in NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, *** < 0.001.

D) Bar graphs showing average expression in CPM of *2700038G22Rik* lncRNA at different timepoints of two 2° MEFs reprogramming RNAseq datasets[23,82] and one 2° TTFs reprogramming single-cell RNAseq

dataset[28]. Error bars represent the SEM (middle graph) and dots represent replicates (right graph). P-values were calculated by two-tailed t-test with unequal variance. * < 0.05, ** < 0.01. TTFs: Tail-Tip Fibroblasts. Also see Extended Data Fig. 6A-B.

E) Bar graph showing average expression in CPM of *SNHG26* lncRNA at different timepoints of a 2° human fibroblasts reprogramming RNAseq dataset[83]. Error bars represent the SEM. P-values were calculated by two-tailed t-test with unequal variance. * < 0.05, ** < 0.01, **** < 0.0001. Also see Extended Data Fig. 6C.

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[84]). TSL: Transcript Support Level.

G) Plots (top) representing expression in TPM of two *Snhg26* lncRNA isoforms in NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Lines represent average expression and shaded areas represent SEM. Pie charts (bottom) represent the proportion of each transcript. P-values were calculated by two-tailed 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). Also see Extended Data Fig. 6E.

Figure 7: *Snhg26* lncRNA enhances the reprogramming route to pluripotency

A) Graphs representing the percentage of mCherry+ cells (i.e. expressing OKS + control (grey) or OKS + *Snhg26* (blue)) or **B)** GFP+ cells (i.e. endogenous *Oct4*) in FBSL + doxycycline medium, or after doxycycline removal in 2iL medium, monitored throughout reprogramming using flow cytometry. Error bars represent SEM (n=3). P-values were calculated by one-tailed t-test with paired samples. n.s. > 0.05, * ≤ 0.05. The data is the same as Fig. 1B-C for OKS + control because experiments were conducted at the same time for OKMS, OKS and OKS + *Snhg26* lncRNA.

C) Growth curves (left) showing average number of mCherry+ cells from the reprogramming cell lines for OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=3) in FBSL + doxycycline medium through the first 5 days of reprogramming. Error bars represent the SEM. P-values were calculated by one-tailed t-test on paired samples. * < 0.05. Bar graph (right) of mean doubling times for mCherry+ cells from the reprogramming cell lines for OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=3) in FBSL + doxycycline medium. Doubling times were calculated between 24 hours and 120 hours of doxycycline

induction. Error bars represent the STD and dots represent replicates. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05. Also see Extended Data Fig. 7C.

D) The density of cells expressing OKS + control or OKS + *Snhg26* lncRNA (mCherry+) or **E)** endogenous *Oct4* (GFP+) in FBSL + doxycycline medium, or after doxycycline removal in 2iL medium, is monitored throughout OKS + control (grey) or OKS + *Snhg26* lncRNA (blue) reprogramming using flow cytometry. Error bars represent the SEM (n=3). P-values were calculated by one-tailed t-test with paired samples. n.s. > 0.05, * ≤ 0.05. Also see Extended Data Fig. 8A.

F) Schematic of the limiting dilution assay where 5, 25, 125 or 500 cells of the OKS + control and OKS + *Snhg26* NPC lines were seeded per well of a 96 well-plate (24 replicates/dilution). Wells negative for iPSCs (GFP+) were counted at D12 and D27 of reprogramming.

G) Graphs representing the mean percentage of GFP-negative wells for the different cell densities at D12 of reprogramming in FBSL + doxycycline medium (n=3) (top) and at D27 of reprogramming in 2iL medium (n=3) (bottom), allowing the use of Poisson distribution. Error bars represent the SEM. Bar graph of average iPSC frequencies in OKS + control and OKS + *Snhg26* conditions at D12 (n=3) and D27 (n=3) of reprogramming. Error bars represent the STD and each dot represents the replicates. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 0.05.

Figure 8: *Snhg26* lncRNA is a developmental gene and is necessary for maintenance of both mouse and human pluripotent state of ESCs

A) Histograms representing the epigenetic status (H3K4me3 = promoter regions; H3K27me3 = silenced region; CpG sites = promoter region)[23] at the genomic loci of *Snhg26*. Tracks represent 2° MEFs (D0) and their reprogrammed counterparts after 16 days of reprogramming (D16) or 2° iPSCs and were visualized with IGV software[114]. ESCs were used as controls for pluripotent cell types.

B) Bar graphs of RNAseq expression of *Snhg26* lncRNA (in TPM) in mouse pre-implantation embryo stages[16] (left), and average single-cell RNAseq expression of *SNHG26* lncRNA (in CPM) in human pre-implantation embryo stages[90] (right). Dots represent outlier cells. P-values were calculated by two-tailed t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001. hESCs: human Embryonic Stem Cells.

C) Bar graph of average RNAseq expression of *SNHG26* lncRNA (in CPM) in primed and *naïve* human ESCs or 8CLCs dataset[95]. Error bars represent the SEM. P-values were calculated by two-tailed 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.

D) Dot plot of mean fold changes in expression for *Snhg26* lncRNA and pluripotency-associated genes, assessed by RT-qPCR 5 days after *Snhg26* knock-down (n=3) in mouse ESCs using the CRISPRi system. Expression is normalized to a control sample with non-targeting sgRNA and to the *Rpl27* housekeeping gene. Error bars represent the STD. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05.

E) Dot plot of mean fold changes in expression for *SNHG26* lncRNA and core, primed or *naïve* pluripotency-associated genes, assessed by RT-qPCR 3 days after *SNHG26* knock-down (n=2) in human ESCs using pooled sgRNAs with the CRISPRi system. Expression is normalized to a control sample with non-targeting sgRNA and to the *RPS13* housekeeping gene. Error bars represent the STD from 6 technical replicates. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, ** < 0.01, **** < 0.0001.

F) Schematic representing the enhancing effect of *Snhg26* lncRNA on iPSC generation and the isoform switch that occurs during reprogramming (left). Comparatively, *Snhg26* knock-down promotes ESC differentiation.

Figure 1

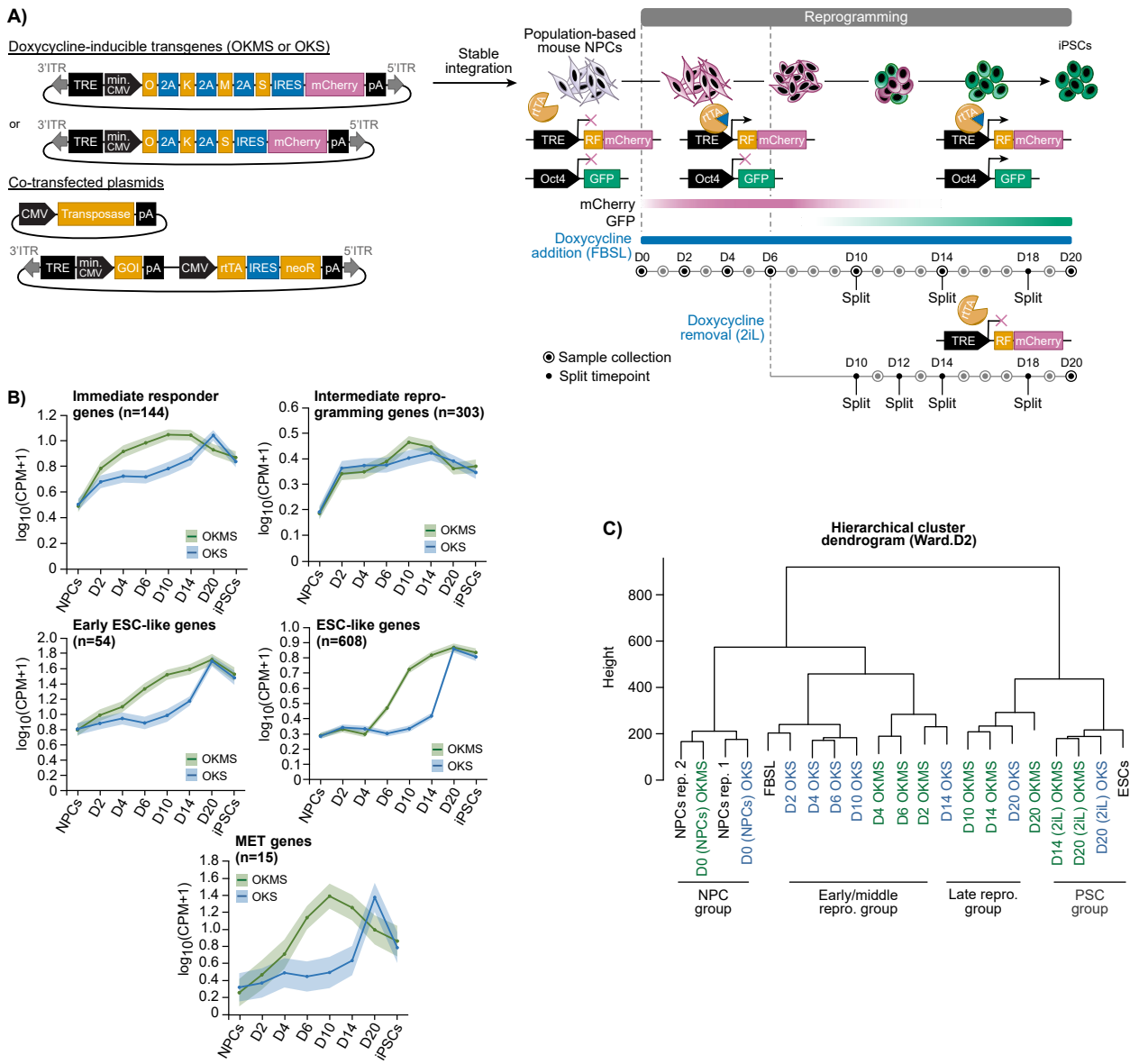

A)
Doxycycline-inducible transgenes (OKMS or OKS)
3'ITR
TRE
min. CMV
O
2A
K
M
S
IRES
mCherry
pA
5'ITR
or
Co-transfected plasmids
CMV
Transposase
GOI
rtTA
neoR
Stable integration
Reprogramming
Population-based mouse NPCs
iPSCs
RF
Oct4
GFP
mCherry
Doxycycline addition (FBSL)
Doxycycline removal (2iL)
Split
Sample collection
Split timepoint
B)
Immediate responder genes (n=144)
Intermediate reprogramming genes (n=303)
Early ESC-like genes (n=54)
ESC-like genes (n=608)
MET genes (n=15)
log10(CPM+1)
OKMS
OKS
NPCs
iPSCs
C)
Hierarchical cluster dendrogram (Ward.D2)
Height
NPCs rep. 2
D0 (NPCs) OKMS
NPCs rep. 1
D0 (NPCs) OKS
FBSL
D2 OKS
D4 OKS
D6 OKS
D10 OKS
D4 OKMS
D6 OKMS
D2 OKMS
D14 OKS
D10 OKMS
D14 OKMS
D20 OKS
D20 OKMS
D14 (2iL) OKMS
D20 (2iL) OKMS
D20 (2iL) OKS
ESCs
NPC group
Early/middle repro. group
Late repro. group
PSC group

Figure 2

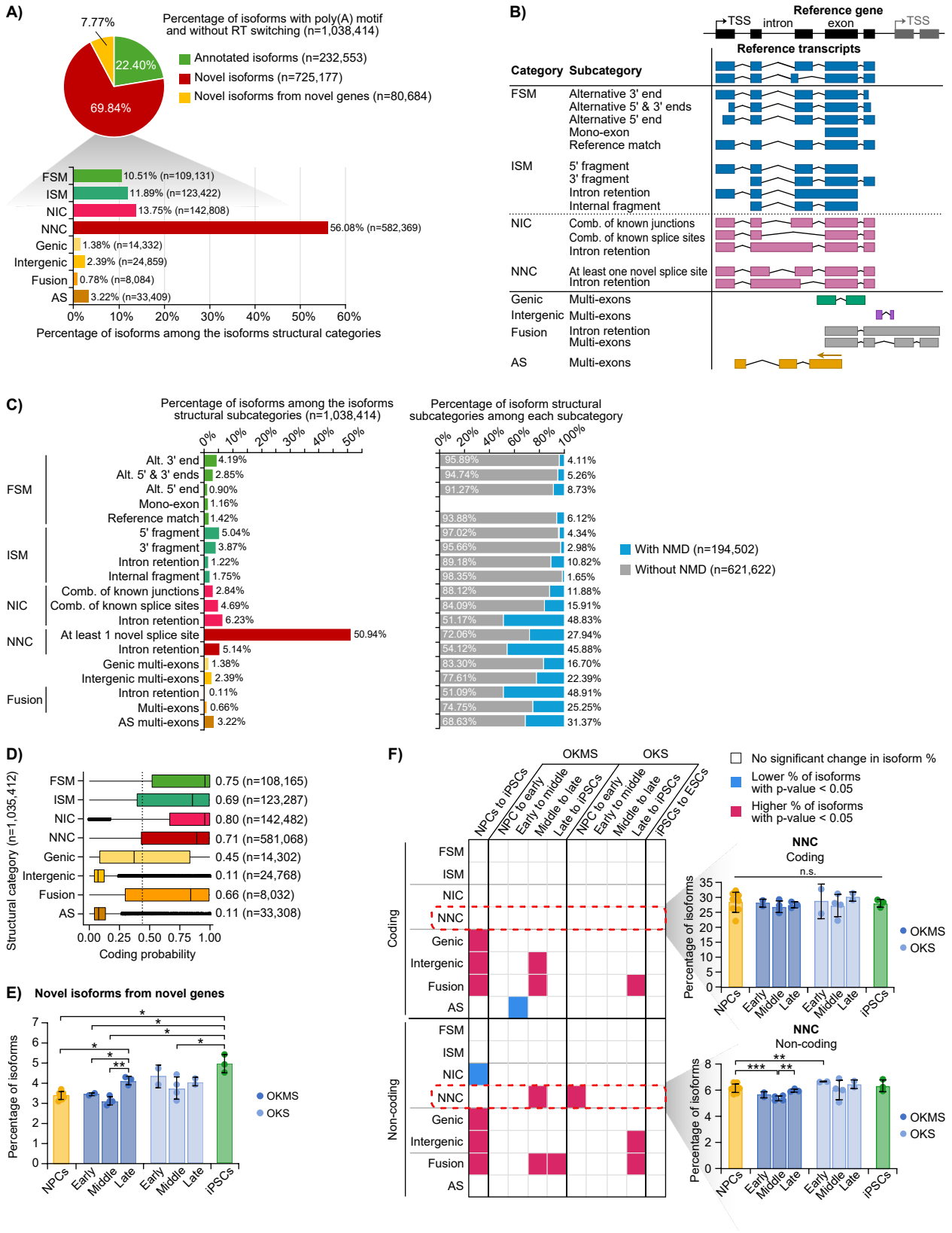

A)
Percentage of isoforms with poly(A) motif and without RT switching (n=1,038,414)
7.77%
22.40%
69.84%
Annotated isoforms (n=232,553)
Novel isoforms (n=725,177)
Novel isoforms from novel genes (n=80,684)
FSM 10.51% (n=109,131)
ISM 11.89% (n=123,422)
NIC 13.75% (n=142,808)
NNC 56.08% (n=582,369)
Genic 1.38% (n=14,332)
Intergenic 2.39% (n=24,859)
Fusion 0.78% (n=8,084)
AS 3.22% (n=33,409)
Percentage of isoforms among the isoforms structural categories
B)
Reference gene
TSS
intron
exon
Reference transcripts
Category
Subcategory
FSM
Alternative 3' end
Alternative 5' & 3' ends
Alternative 5' end
Mono-exon
Reference match
ISM
5' fragment
3' fragment
Intron retention
Internal fragment
NIC
Comb. of known junctions
Comb. of known splice sites
Intron retention
NNC
At least one novel splice site
Intron retention
Genic
Multi-exons
Intergenic
Multi-exons
Fusion
Intron retention
Multi-exons
AS
Multi-exons
C)
Percentage of isoforms among the isoforms structural subcategories (n=1,038,414)
Percentage of isoform structural subcategories among each subcategory
Alt. 3' end 4.19%
Alt. 5' & 3' ends 2.85%
Alt. 5' end 0.90%
Mono-exon 1.16%
Reference match 1.42%
5' fragment 5.04%
3' fragment 3.87%
Intron retention 1.22%
Internal fragment 1.75%
Comb. of known junctions 2.84%
Comb. of known splice sites 4.69%
Intron retention 6.23%
At least 1 novel splice site 50.94%
Intron retention 5.14%
Genic multi-exons 1.38%
Intergenic multi-exons 2.39%
Intron retention 0.11%
Multi-exons 0.66%
AS multi-exons 3.22%
With NMD (n=194,502)
Without NMD (n=621,622)
D)
Structural category (n=1,035,412)
FSM 0.75 (n=108,165)
ISM 0.69 (n=123,287)
NIC 0.80 (n=142,482)
NNC 0.71 (n=581,068)
Genic 0.45 (n=14,302)
Intergenic 0.11 (n=24,768)
Fusion 0.66 (n=8,032)
AS 0.11 (n=33,308)
Coding probability
E) Novel isoforms from novel genes
Percentage of isoforms
OKMS
OKS
NPCs
Early
Middle
Late
iPSCs
F)
OKMS
OKS
NPCs to iPSCs
NPC to early
Early to middle
Middle to late
Late to iPSCs
NPC to early
Early to middle
Middle to late
Late to iPSCs
iPSCs to ESCs
No significant change in isoform %
Lower % of isoforms with p-value < 0.05
Higher % of isoforms with p-value < 0.05
Coding
Non-coding
FSM
ISM
NIC
NNC
Genic
Intergenic
Fusion
AS
NNC
Coding
n.s.
NNC
Non-coding

Figure 3

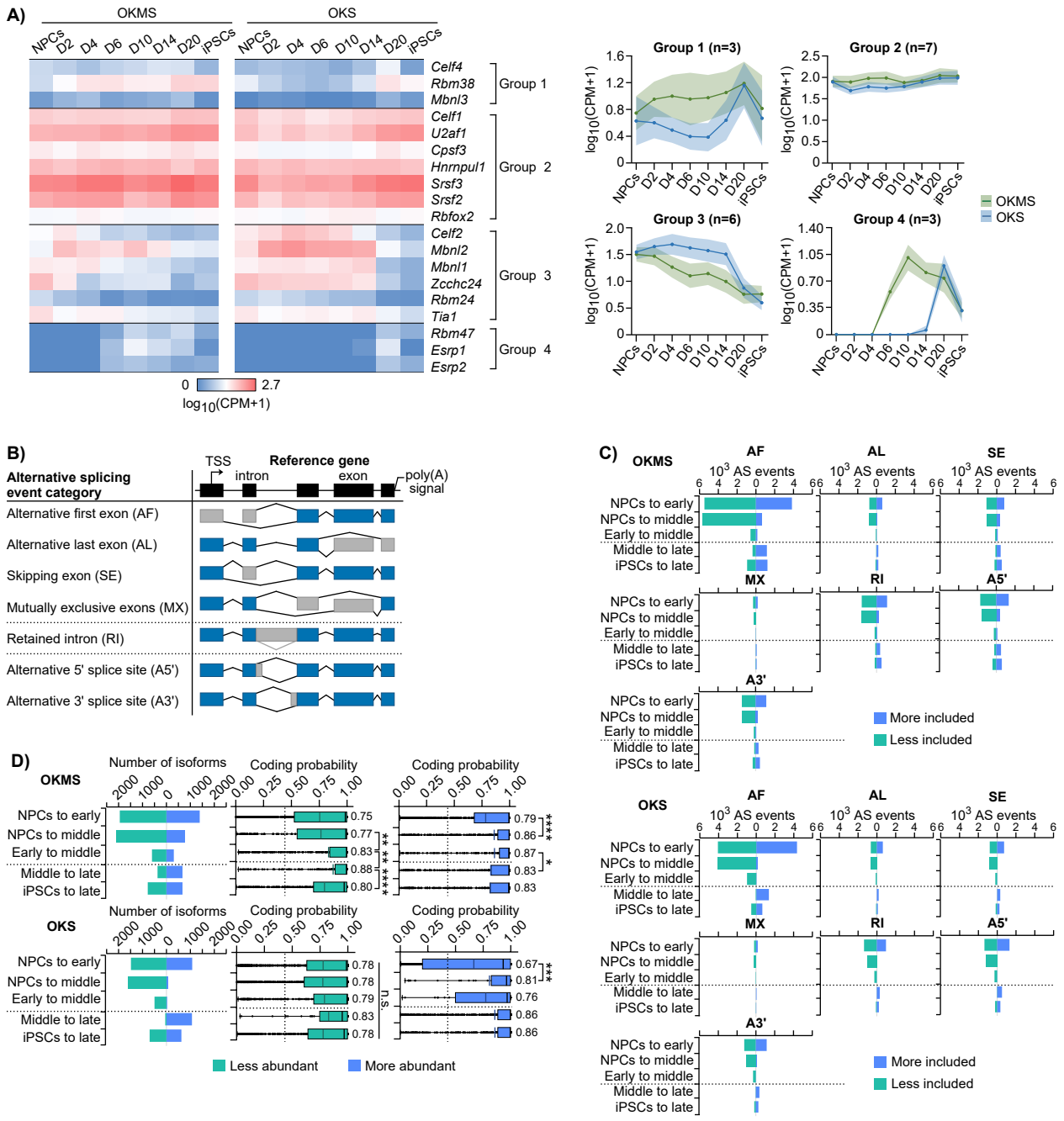

A)
OKMS
OKS
NPCs D2 D4 D6 D10 D14 D20 iPSCs
Celf4
Rbm38
Mbnl3
Celf1
U2af1
Cpsf3
Hnrnpul1
Srsf3
Srsf2
Rbfox2
Celf2
Mbnl2
Mbnl1
Zcchc24
Rbm24
Tia1
Rbm47
Esrp1
Esrp2
Group 1
Group 2
Group 3
Group 4
0
2.7
log10(CPM+1)
Group 1 (n=3)
Group 2 (n=7)
Group 3 (n=6)
Group 4 (n=3)
OKMS
OKS
B)
Alternative splicing 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
OKS
AF
AL
SE
MX
RI
A5'
A3'
10^3 AS events
NPCs to early
NPCs to middle
Early to middle
Middle to late
iPSCs to late
More included
Less included
D)
OKMS
OKS
Number of isoforms
Coding probability
Less abundant
More abundant

Figure 4

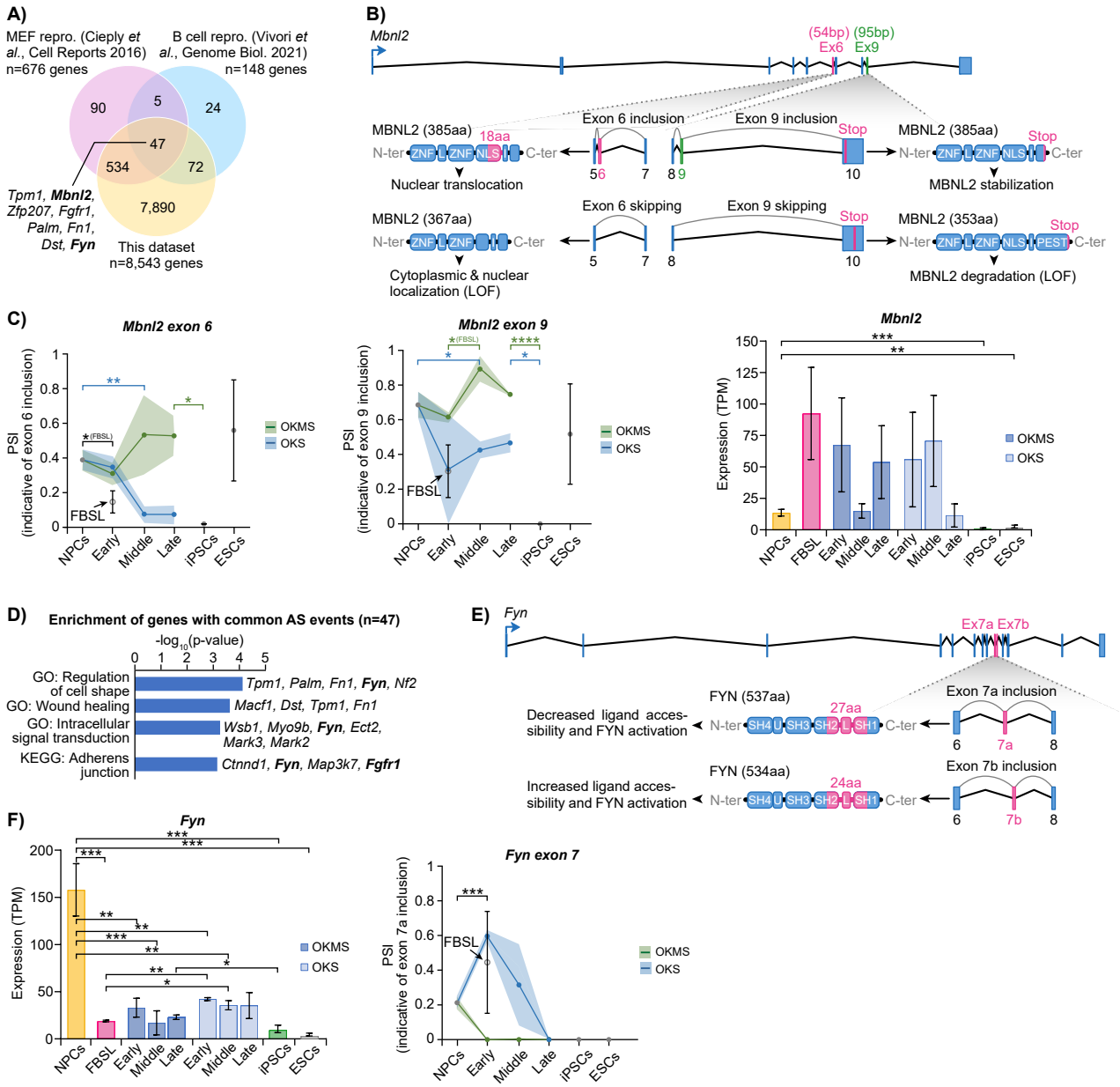

A)
MEF repro. (Cieply et al., Cell Reports 2016) n=676 genes
B cell repro. (Vivori et al., Genome Biol. 2021) n=148 genes
90
5
24
47
534
72
7,890
Tpm1, Mbnl2, Zfp207, Fgfr1, Palm, Fn1, Dst, Fyn
This dataset n=8,543 genes
B)
Mbnl2
(54bp) (95bp)
Ex6 Ex9
MBNL2 (385aa)
18aa
Exon 6 inclusion
Exon 9 inclusion
Stop
MBNL2 (385aa)
Nuclear translocation
MBNL2 stabilization
MBNL2 (367aa)
Exon 6 skipping
Exon 9 skipping
MBNL2 (353aa)
Cytoplasmic & nuclear localization (LOF)
MBNL2 degradation (LOF)
C)
Mbnl2 exon 6
PSI (indicative of exon 6 inclusion)
Mbnl2 exon 9
PSI (indicative of exon 9 inclusion)
Mbnl2
Expression (TPM)
OKMS
OKS
FBSL
NPCs
Early
Middle
Late
iPSCs
ESCs
D)
Enrichment of genes with common AS events (n=47)
-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
E)
Fyn
Ex7a Ex7b
FYN (537aa)
27aa
Exon 7a inclusion
Decreased ligand accessibility and FYN activation
FYN (534aa)
24aa
Exon 7b inclusion
Increased ligand accessibility and FYN activation
F)
Fyn
Fyn exon 7
PSI (indicative of exon 7a inclusion)

Figure 5

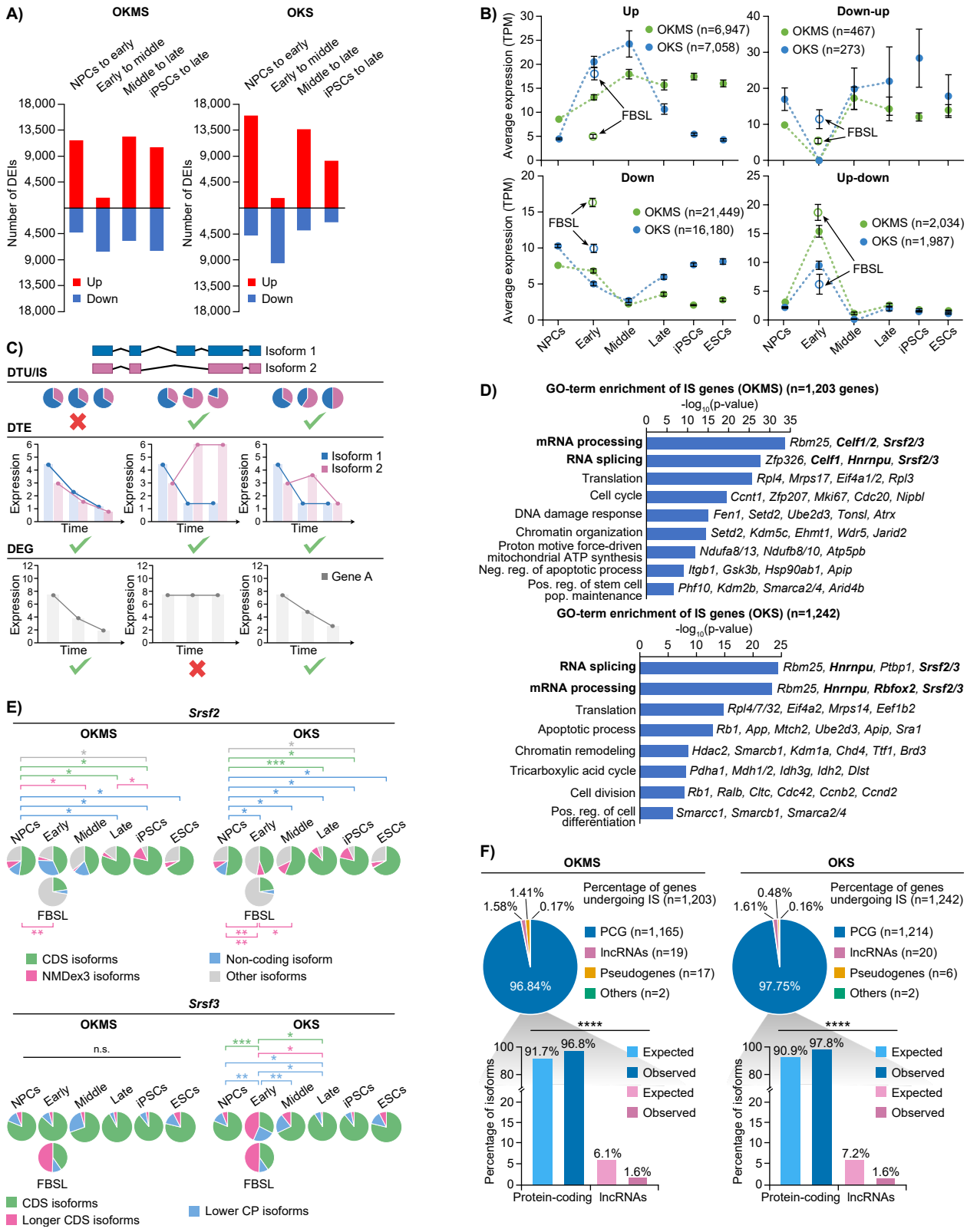

A)
OKMS
OKS
NPCs to early
Early to middle
Middle to late
iPSCs to late
Number of DEIs
Up
Down
B)
Up
Down-up
Down
Up-down
OKMS (n=6,947)
OKS (n=7,058)
OKMS (n=467)
OKS (n=273)
OKMS (n=21,449)
OKS (n=16,180)
OKMS (n=2,034)
OKS (n=1,987)
FBSL
Average expression (TPM)
NPCs
Early
Middle
Late
iPSCs
ESCs
C)
DTU/IS
Isoform 1
Isoform 2
DTE
DEG
Expression
Time
Gene A
D)
GO-term enrichment of IS genes (OKMS) (n=1,203 genes)
-log10(p-value)
mRNA processing
RNA splicing
Translation
Cell cycle
DNA damage response
Chromatin organization
Proton motive force-driven mitochondrial ATP synthesis
Neg. reg. of apoptotic process
Pos. reg. of stem cell pop. maintenance
Rbm25, Celf1/2, Srsf2/3
Zfp326, Celf1, Hnrnpu, Srsf2/3
Rpl4, Mrps17, Eif4a1/2, Rpl3
Ccnt1, Zfp207, Mki67, Cdc20, Nipbl
Fen1, Setd2, Ube2d3, Tonsl, Atrx
Setd2, Kdm5c, Ehmt1, Wdr5, Jarid2
Ndufa8/13, Ndufb8/10, Atp5pb
Itgb1, Gsk3b, Hsp90ab1, Apip
Phf10, Kdm2b, Smarca2/4, Arid4b
GO-term enrichment of IS genes (OKS) (n=1,242)
-log10(p-value)
RNA splicing
mRNA processing
Translation
Apoptotic process
Chromatin remodeling
Tricarboxylic acid cycle
Cell division
Pos. reg. of cell differentiation
Rbm25, Hnrnpu, Ptbp1, Srsf2/3
Rbm25, Hnrnpu, Rbfox2, Srsf2/3
Rpl4/7/32, Eif4a2, Mrps14, Eef1b2
Rb1, App, Mtch2, Ube2d3, Apip, Sra1
Hdac2, Smarcb1, Kdm1a, Chd4, Ttf1, Brd3
Pdha1, Mdh1/2, Idh3g, Idh2, Dlst
Rb1, Ralb, Cltc, Cdc42, Ccnb2, Ccnd2
Smarcc1, Smarcb1, Smarca2/4
E)
Srsf2
Srsf3
n.s.
CDS isoforms
NMDex3 isoforms
Non-coding isoform
Other isoforms
Longer CDS isoforms
Lower CP isoforms
F)
Percentage of genes undergoing IS (n=1,203)
Percentage of genes undergoing IS (n=1,242)
PCG (n=1,165)
lncRNAs (n=19)
Pseudogenes (n=17)
Others (n=2)
PCG (n=1,214)
lncRNAs (n=20)
Pseudogenes (n=6)
Others (n=2)
96.84%
97.75%
Expected
Observed
Percentage of isoforms
Protein-coding
lncRNAs

Figure 6

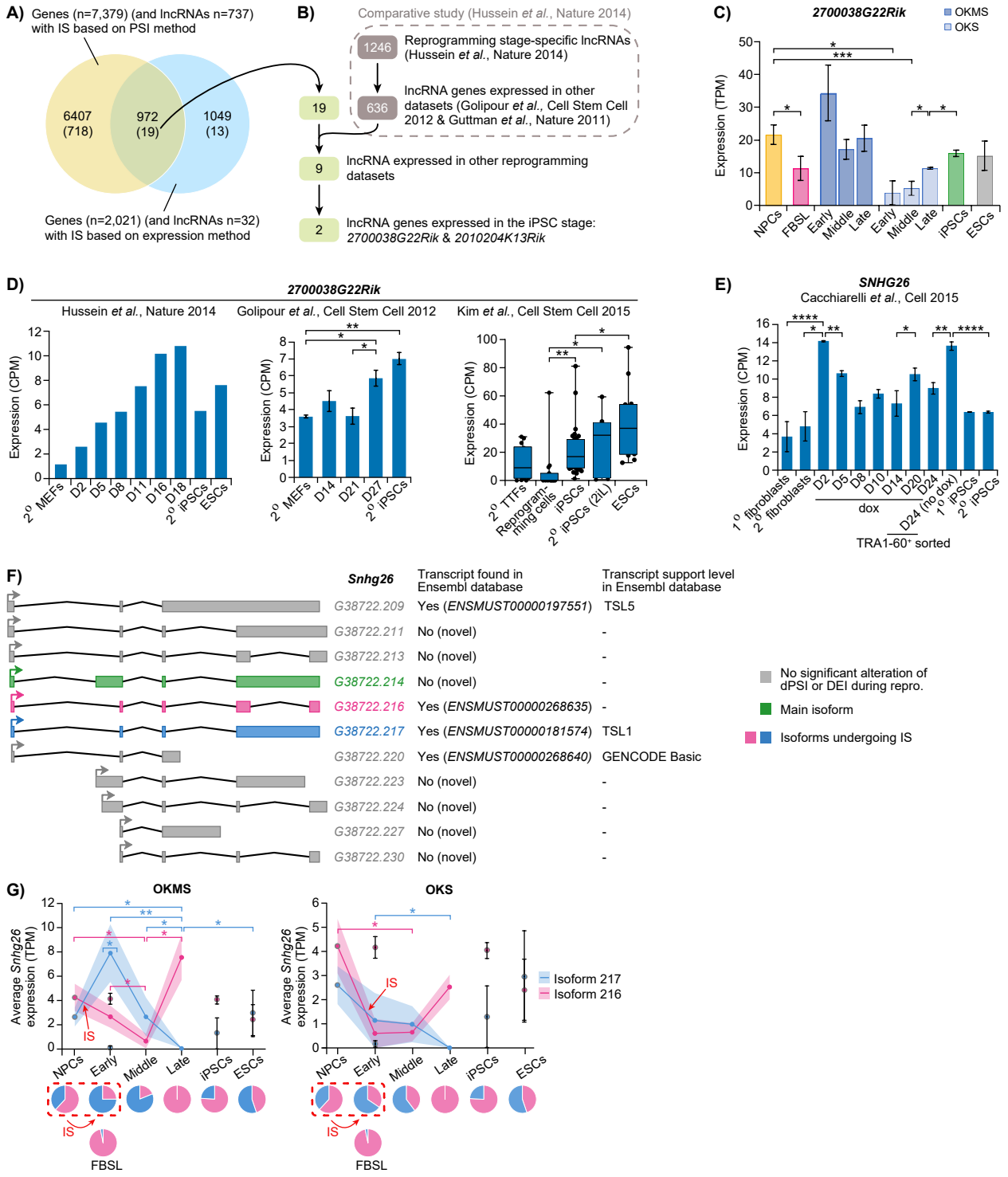

A)
Genes (n=7,379) (and lncRNAs n=737) with IS based on PSI method
6407 (718)
972 (19)
1049 (13)
Genes (n=2,021) (and lncRNAs n=32) with IS based on expression method
B)
Comparative study (Hussein et al., Nature 2014)
1246
Reprogramming stage-specific lncRNAs (Hussein et al., Nature 2014)
636
lncRNA genes expressed in other datasets (Golipour et al., Cell Stem Cell 2012 & Guttman et al., Nature 2011)
19
9
lncRNA expressed in other reprogramming datasets
2
lncRNA genes expressed in the iPSC stage: 2700038G22Rik & 2010204K13Rik
C)
2700038G22Rik
OKMS
OKS
Expression (TPM)
NPCs
FBSL
Early
Middle
Late
iPSCs
ESCs
D)
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
2° iPSCs (2iL)
E)
SNHG26
Cacchiarelli et al., Cell 2015
Expression (CPM)
1° fibroblasts
2° fibroblasts
D10
D20
D24
D24 (no dox)
1° iPSCs
dox
TRA1-60+ sorted
F)
Snhg26
Transcript found in Ensembl database
Transcript support level in Ensembl database
G38722.209 Yes (ENSMUST00000197551) TSL5
G38722.211 No (novel) -
G38722.213 No (novel) -
G38722.214 No (novel) -
G38722.216 Yes (ENSMUST00000268635) -
G38722.217 Yes (ENSMUST00000181574) TSL1
G38722.220 Yes (ENSMUST00000268640) GENCODE Basic
G38722.223 No (novel) -
G38722.224 No (novel) -
G38722.227 No (novel) -
G38722.230 No (novel) -
No significant alteration of dPSI or DEI during repro.
Main isoform
Isoforms undergoing IS
G)
OKMS
OKS
Average Snhg26 expression (TPM)
IS
Isoform 217
Isoform 216
FBSL

Figure 7

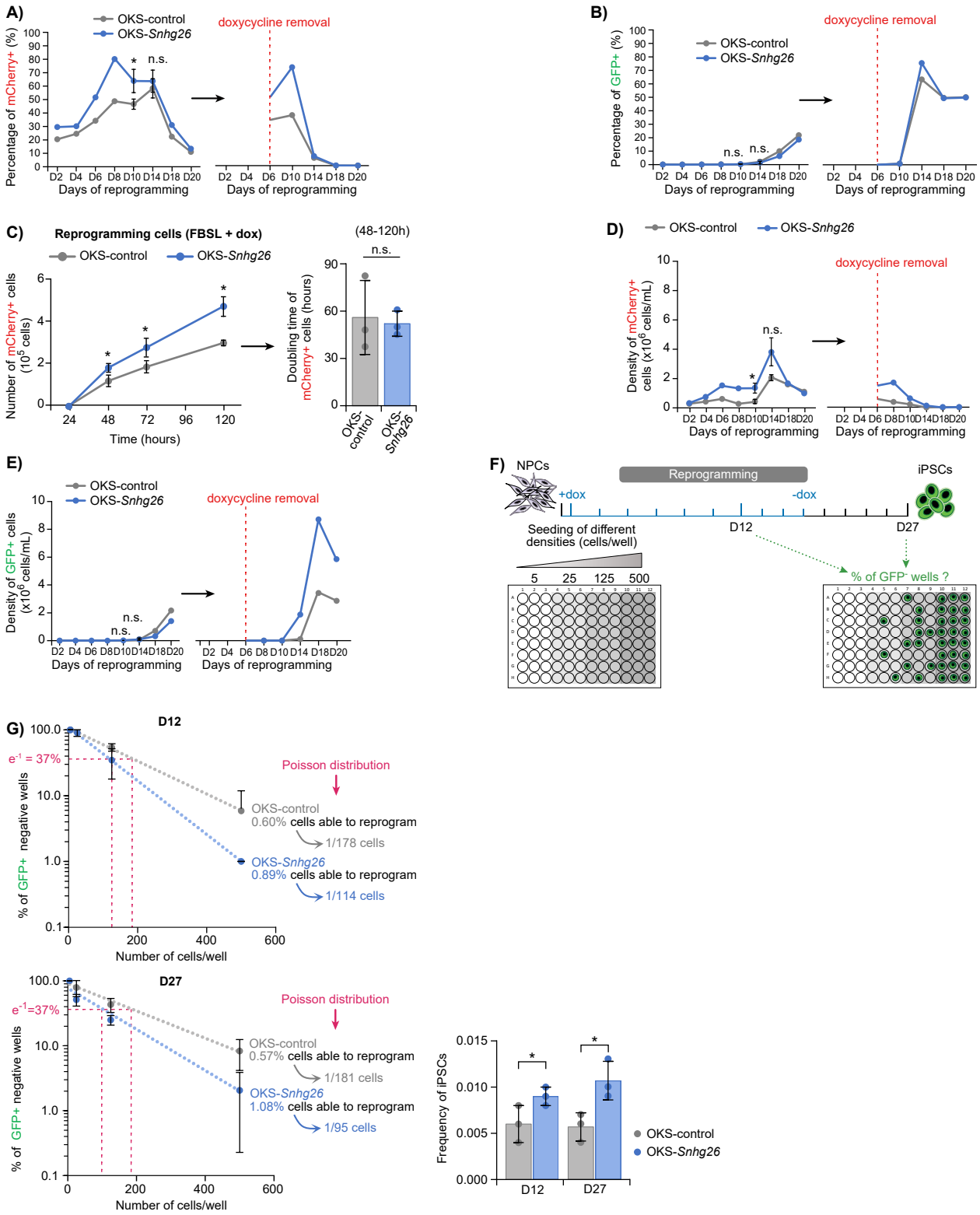
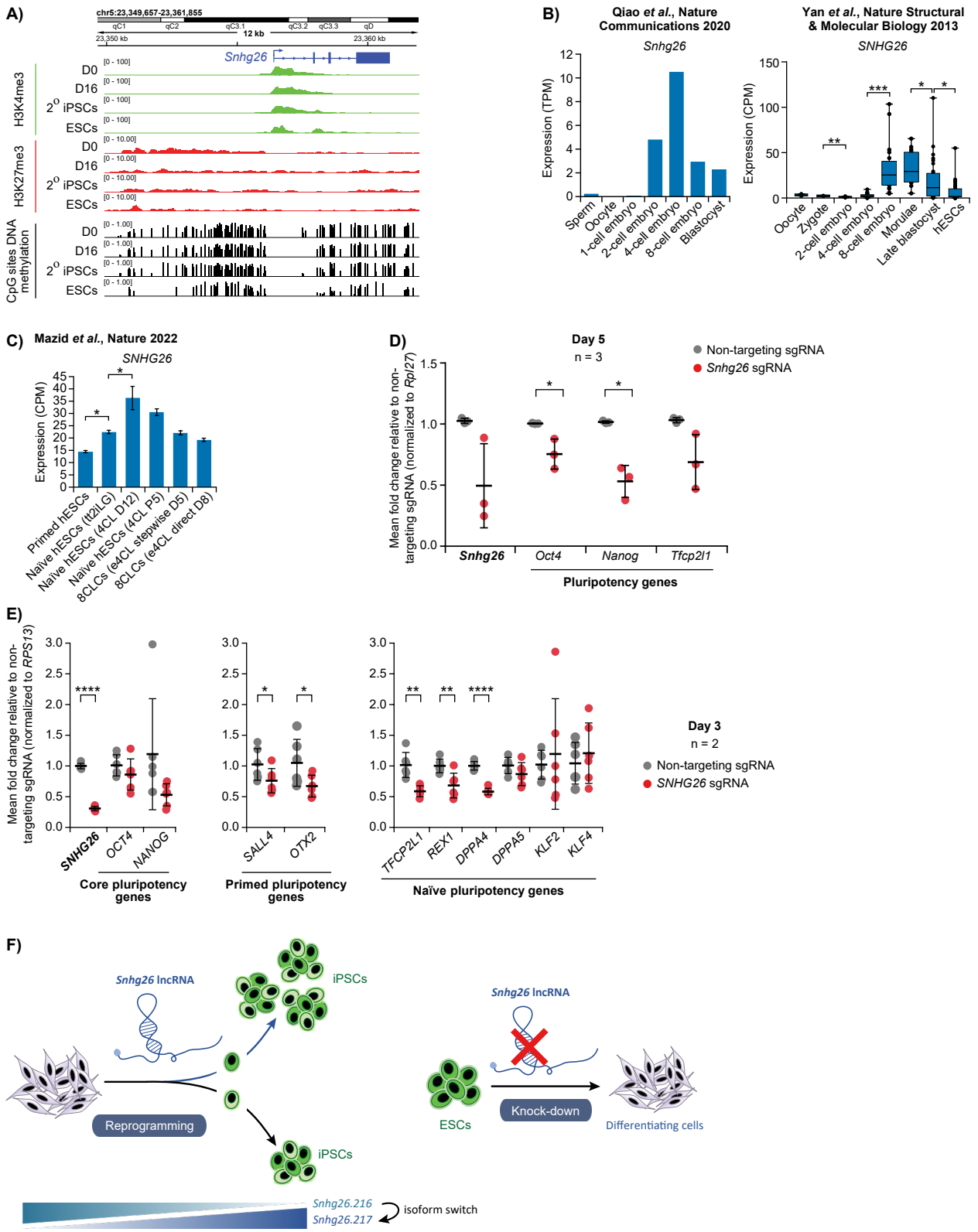

A)
OKS-control
OKS-Snhg26
doxycycline removal
Percentage of mCherry+ (%)
Days of reprogramming
B)
Percentage of GFP+ (%)
C)
Reprogramming cells (FBSL + dox)
Number of mCherry+ cells (10^5 cells)
Time (hours)
(48-120h)
Doubling time of mCherry+ cells (hours)
D)
Density of mCherry+ cells (x10^6 cells/mL)
E)
Density of GFP+ cells (x10^6 cells/mL)
F)
NPCs
Reprogramming
iPSCs
+dox
-dox
Seeding of different densities (cells/well)
D12
D27
% of GFP- wells ?
G)
D12
% of GFP+ negative wells
Poisson distribution
OKS-control
0.60% cells able to reprogram
1/178 cells
OKS-Snhg26
0.89% cells able to reprogram
1/114 cells
Number of cells/well
D27
0.57% cells able to reprogram
1/181 cells
1.08% cells able to reprogram
1/95 cells
Frequency of iPSCs

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)
Qiao et al., Nature Communications 2020
Snhg26
Expression (TPM)
Sperm
Oocyte
1-cell embryo
2-cell embryo
4-cell embryo
8-cell embryo
Blastocyst
Yan et al., Nature Structural & Molecular Biology 2013
SNHG26
Expression (CPM)
Oocyte
Zygote
2-cell embryo
4-cell embryo
8-cell embryo
Morulae
Late blastocyst
hESCs
C)
Mazid et al., Nature 2022
SNHG26
Expression (CPM)
Primed hESCs
Naïve hESCs (t2iLG)
Naïve hESCs (4CL D12)
Naïve hESCs (4CL P5)
Naïve hESCs (4CL stepwise D5)
8CLCs (e4CL direct D8)
D)
Day 5
n = 3
Non-targeting sgRNA
Snhg26 sgRNA
Mean fold change relative to non-targeting sgRNA (normalized to Rpl27)
Snhg26
Oct4
Nanog
Tfcp2l1
Pluripotency genes
E)
Mean fold change relative to non-targeting sgRNA (normalized to RPS13)
SNHG26
OCT4
NANOG
Core pluripotency genes
SALL4
OTX2
Primed pluripotency genes
TFCP2L1
REX1
DPPA4
DPPA5
KLF2
KLF4
Naïve pluripotency genes
Day 3
n = 2
Non-targeting sgRNA
SNHG26 sgRNA
F)
Snhg26 lncRNA
iPSCs
Reprogramming
Snhg26 lncRNA
ESCs
Knock-down
Differentiating cells
iPSCs
Snhg26.216
Snhg26.217
isoform switch

Extended Data Figure legends

Extended Data Figure 1: Designing an efficient reprogramming system, related to Figure 1

A) Representative images in brightfield (upper track) or fluorescence (middle and bottom tracks) microscopy of the NPCs (D0), reprogramming cells at different timepoints, and final iPSCs (D20 2iL) in the OKMS (top) and OKS (bottom) reprogramming systems. dox: doxycycline. Scalebar = 50 µm.

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

Extended Data Figure 2: Designing an efficient reprogramming system, related to Figure 1

A) Plot depicting the logged expression in CPM of epithelial-to-mesenchymal transition (EMT) (n=20) genes in each timepoint of the OKMS (green) or OKS (blue) reprogramming. Lines represent the average expression value at each timepoint and the shaded area represents the SEM.

B) Bar graphs depicting the total number of genes that are significantly up- (red) or down- (blue) regulated between the different timepoints of OKMS (left) and OKS (right) reprogramming. Significance based on FDR ≤ 0.05 and log2(fold change) ≥ 2 for up-regulated genes or ≤ 0.5 for down-regulated genes. DEGs: Differentially Expressed Genes.

C) Hierarchical clustering on whole transcriptome of the OKMS (top) or OKS (bottom) samples (clustering method: Ward.D2). The data is the same in OKMS and OKS and in Fig. 1F for NPC, FBSL and PSC samples because those samples were in common for OKMS and OKS.

Extended Data Figure 3: lrRNAseq reveals unprecedented isoform diversity and that non-coding isoforms are exclusively regulated during reprogramming, related to Figure 2

A) Schematic overview of the experimental and analysis pipeline used to generate full-length transcript annotations of OKMS and OKS reprogramming samples. Samples from two different runs of lrRNAseq were used. First run was performed using poly(A)+ RNA enrichment followed by direct cDNA native barcoding and loading on R9 flow cells, whereas the second run was generated using RNA directly retrotranscribed and then PCR-amplified before loading on R10 flow cells. The sequencing was performed using either the MinION or PromethION 2 Solo devices from Oxford Nanopore Technologies (ONT).

Reads were processed for quality control and then aligned to the mouse transcriptome. Differential gene expression was performed using EdgeR, whereas isoform classification was done with SQANTI 3 software, splicing analyses were run with SUPPA 2 software, and differential isoform usage was analyzed using IsoQuant. gDNA: genomic DNA; RT: Reverse Transcription; cDNA: complementary DNA; PCR: Polymerase Chain Reaction; AS: Alternative Splicing; adj. p-value: adjusted p-value; lncRNAs: long non-coding RNAs.

B) Pie chart representing the proportion of isoforms with (blue) or without (yellow) a poly(A) motif among the total detected transcripts of the generated dataset.

C) Pie chart and bar graph representing the proportion of coding (orange) and non-coding (blue) isoforms detected in the filtered dataset (top), or separated by isoform structural category (bottom). Interg.: Intergenic.

D) Bar graphs representing the mean percentage of genic, intergenic, fusion and antisense isoforms detected in the NPCs (n=8), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) of the filtered dataset. Error bars represent the STD and each dot represents the replicates. P-values were calculated by two-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01.

Extended Data Figure 4: Alternative splicing follows a biphasic pattern during reprogramming, related to Figure 3

A) Number of significant alternative splicing events more (blue) or less (turquoise) included at each timepoint transition of OKMS (top) or OKS (bottom) reprogramming. Significance based on a $\Delta PSI \geq 0.1$ (more included) or $\Delta PSI \leq 0.1$ (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.

B) Number of significant alternatively spliced isoforms more (blue) or less (turquoise) abundant at each timepoint transition of OKMS (top) or OKS (bottom) reprogramming. Significance based on a $\Delta PSI \geq 0.1$ (more abundant) or $\Delta PSI \leq 0.1$ (less abundant) 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.

Extended Data Figure 5: Isoform switching selects against lncRNAs during early phase reprogramming, related to Figure 5

A) Bar graph depicting the total number of isoforms that are significantly up- (red) or down- (blue) regulated between the different phases of OKMS and OKS reprogramming among the total isoforms (n=154,355) detected in the filtered 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) Graphs depicting the total number of coding (left) and non-coding (right) isoforms that are significantly up- (red) or down- (blue) regulated between the different phases of OKMS (left) and OKS (right) 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) Bar graph representing the mean expression in TPM of *Srsf2* gene in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, *** < 0.001.

D) Schematic depicting the transcripts of the *Srs2* gene detected in our dataset, as visualized with IGV software[114].

E) Plots representing the expression in TPM of *Srsf2* isoforms in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Lines represent the average expression value at each timepoint and the shaded area represents the SEM. P-values were calculated by one-tailed 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).

F) Bar graph representing the mean expression in TPM of *Srsf3* gene in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Error bars represent the SEM. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

G) Schematic (left) depicting the transcripts of the *Srs3* gene detected in our dataset, as visualized with IGV software[114], and highlighting the differences in CDS leading to a loss of SRSF3 function. RRM: RNA Recognition Motif domain; RS: Arginine-Serine rich domain. Bar graph (right) showing the coding probability of the *Srsf3* transcript categories. Horizontal dotted line represents the coding probability

optimum cut-off indicating coding transcripts in mouse, according to the CPAT tool[118]. P-values were calculated by two-tailed t-test with unequal variance. **** < 0.0001.

H) Plots representing the expression in TPM of *Srsf3* isoforms in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Lines represent the average expression value at each timepoint and the shaded area represents the SEM. P-values were calculated by one-tailed 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).

Extended Data Figure 6: Identification of a candidate non-coding transcript regulating the pluripotent state, related to Figure 6

A) and B) Bar graphs representing the average expression in CPM of the 8 candidate lncRNAs aside *2700038G22Rik* at different timepoints of two 2° MEFs reprogramming RNAseq datasets[23,82]. Error bars represent the SEM. P-values were calculated by two-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 0.05, ** < 0.01.

C) Bar graph representing the average expression in CPM of *PAGE4* lncRNA at different timepoints of a 2° human fibroblasts reprogramming RNAseq dataset[83]. Error bars represent the SEM. P-values were calculated by two-tailed t-test with unequal variance. n.s. ≥ 0.05.

D) Plots representing the expression in TPM of the main *Snhg26* lncRNA isoform in the NPCs (n=8), FBSL samples (n=3), during early (n=2), middle (n=4) or late (n=3 for OKMS and n=2 for OKS) phases of reprogramming, or in the iPSCs (n=3) and ESCs (n=3) of the filtered dataset. Lines represent the average expression value at each timepoint and the shaded area represents the SEM. P-values were calculated by two-tailed t-test with unequal variance. n.s. ≥ 0.05, * < 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).

E) Plots depicting the PSI values of *Snhg26* transcripts abundance in the same samples of OKMS (top) or OKS (bottom) reprogramming. Lines represent the average PSI value at each timepoint and the shaded area represents the SEM. P-values were calculated by two-tailed 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).

Extended Data Figure 7: *Snhg26* lncRNA enhances the reprogramming route to pluripotency, related to Figure 7

A) Representative images in brightfield (upper track) or fluorescence (middle and bottom tracks) microscopy of the NPCs (D0), reprogramming cells at different timepoints, and final iPSCs (D20 2iL) in the OKS + *Snhg26* lncRNA reprogramming system. Scalebar = 50 µm. Control condition is presented in Extended Data Fig. 1A.

B) Representative flow cytometry data allowing quantification of the percentages of reprogramming cells (mCherry+) and reprogrammed iPSCs (GFP+) throughout the reprogramming with OKS + *Snhg26* lncRNA. Control condition is presented in Extended Data Fig. 1B.

C) Growth curves showing the average number of total cells from the reprogramming cell lines for the OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=4) conditions in FBSL + doxycycline medium through the first 5 days of reprogramming (left). Error bars represent the SEM. P-values were calculated by one-tailed t-test on paired samples. * < 0.05. Bar graphs representing the mean doubling time of the total cell populations from the OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=4) conditions in FBSL + doxycycline medium (right). Doubling times were calculated between 24 hours and 72 hours of doxycycline induction. Error bars represent the STD and dots represent replicates. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05.

D) Growth curves (left) showing the average number of mCherry+ cells from the reprogramming cell lines for the OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=4) conditions in NPC + doxycycline medium through the first 5 days of reprogramming (left). Error bars represent the SEM. P-values were calculated by one-tailed t-test with unequal variance. * < 0.05. Bar graph (right) representing the mean doubling time of the mCherry+ cells from the reprogramming cell lines for the OKS + control transcript (n=3) and OKS + *Snhg26* lncRNA (n=4) conditions in NPC + doxycycline medium (right). Doubling time were calculated between 48 hours and 120 hours of doxycycline induction. Error bars represent the STD and dot represents replicates. P-values were calculated by one-tailed t-test with unequal variance. n.s. ≥ 0.05.

E) Growth curves showing the number of cells from the ESC lines for the wild-type (n=1) and *Snhg26* lncRNA overexpression (n=1) conditions in 2iL + doxycycline medium through the first 2 days of reprogramming.

Extended Data Figure 8: *Snhg26* lncRNA enhances the reprogramming route to pluripotency, related to Figure 7

A) The density of cells not reprogramming in FBSL + doxycycline medium, or after doxycycline removal in 2iL medium, is monitored throughout OKS + control (grey) or OKS + *Snhg26* lncRNA (blue) reprogramming using flow cytometry. Error bars represent the SEM (n=3). P-values were calculated by one-tailed t-test with paired samples. n.s. > 0.05.

B) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D4, for both OKMS + control (grey, n=1) or OKMS + *Snhg26* lncRNA (blue, n=1) reprogramming systems using stable NPC lines, as measured by flow cytometry.

C) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D4, for both OSKM + control (grey, n=1) or OSKM + *Snhg26* lncRNA (blue, n=1) reprogramming systems using stable NPC lines, as measured by flow cytometry.

D) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D12, for both OKS + control (grey, n=1) or OKS + *Snhg26* lncRNA (blue, n=1) reprogramming systems using transient transfections, as measured by flow cytometry.

E) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D4, for both OKMS + control (grey, n=1) or OKMS + *Snhg26* lncRNA (blue, n=1) reprogramming systems using transient transfections, as measured by flow cytometry.

F) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D17, for both OSKM + control (grey, n=1) or OSKM + *Snhg26* lncRNA (blue, n=1) reprogramming systems using transient transfections, as measured by flow cytometry.

G) Density of GFP+ cells in FBSL + doxycycline medium, or in 2iL medium after doxycycline removal at D8, for both OSKM + control (grey, n=1) or OSKM + *Snhg26* lncRNA (blue, n=1) reprogramming systems using ESC line-derived NPCs, as measured by flow cytometry.

H) Representative images of E8-8.5 iPSC-embryo chimeras for assessment of an OKS + control (n=15) or OKS + *Snhg26* lncRNA (n=1) iPSC clone pluripotency after injection into wild-type tdTomato+ embryos. The contribution of iPSC-derived cells was visualized as negative cells/area in tdTomato+ iPSC-embryo chimeras.

Extended Data Figure 9: *Snhg26* lncRNA is a developmental gene and is necessary for maintenance of both mouse and human pluripotent state of ESCs, related to Figure 8

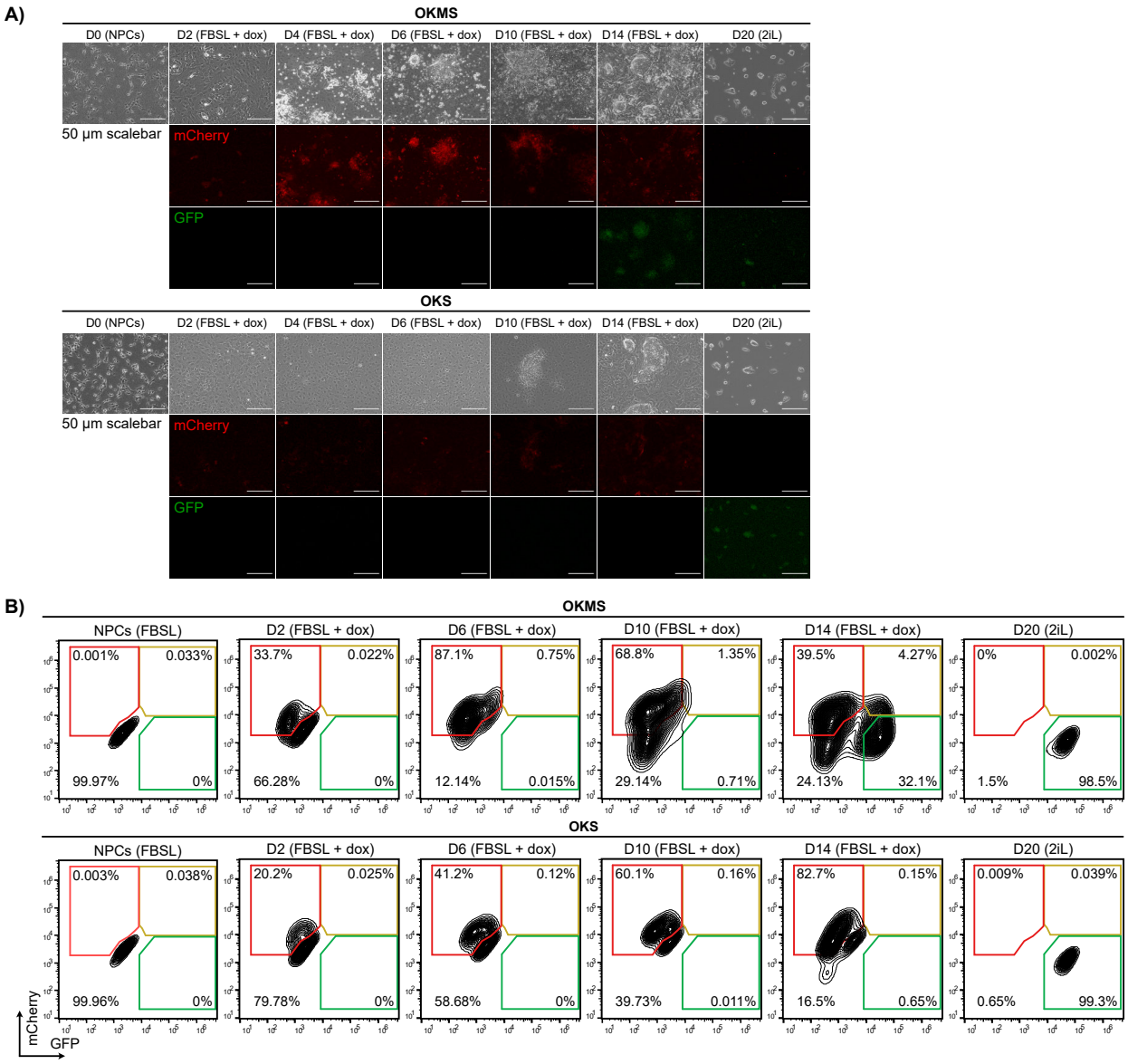
A) Bar graphs representing the expression in CPM of *Snhg26* or *SNHG26* lncRNAs at different timepoints of *in vitro* differentiation into cortical neurons RNAseq datasets from mouse[92] (left) or human[91] (right) ESCs.

B) Graph representing the mean percentage of *Snhg26* lncRNA in the nuclear or cytoplasmic compartments after mouse ESC fractionation (n=3), normalized to total RNA, as assessed by RT-qPCR. Expression levels of other mRNAs and lncRNAs are shown as controls. Error bars represent the STD and each dot represents the replicates. P-values were calculated by two-tailed t-test with unequal variance. **** < 0.0001. mRNAs: messenger RNAs.

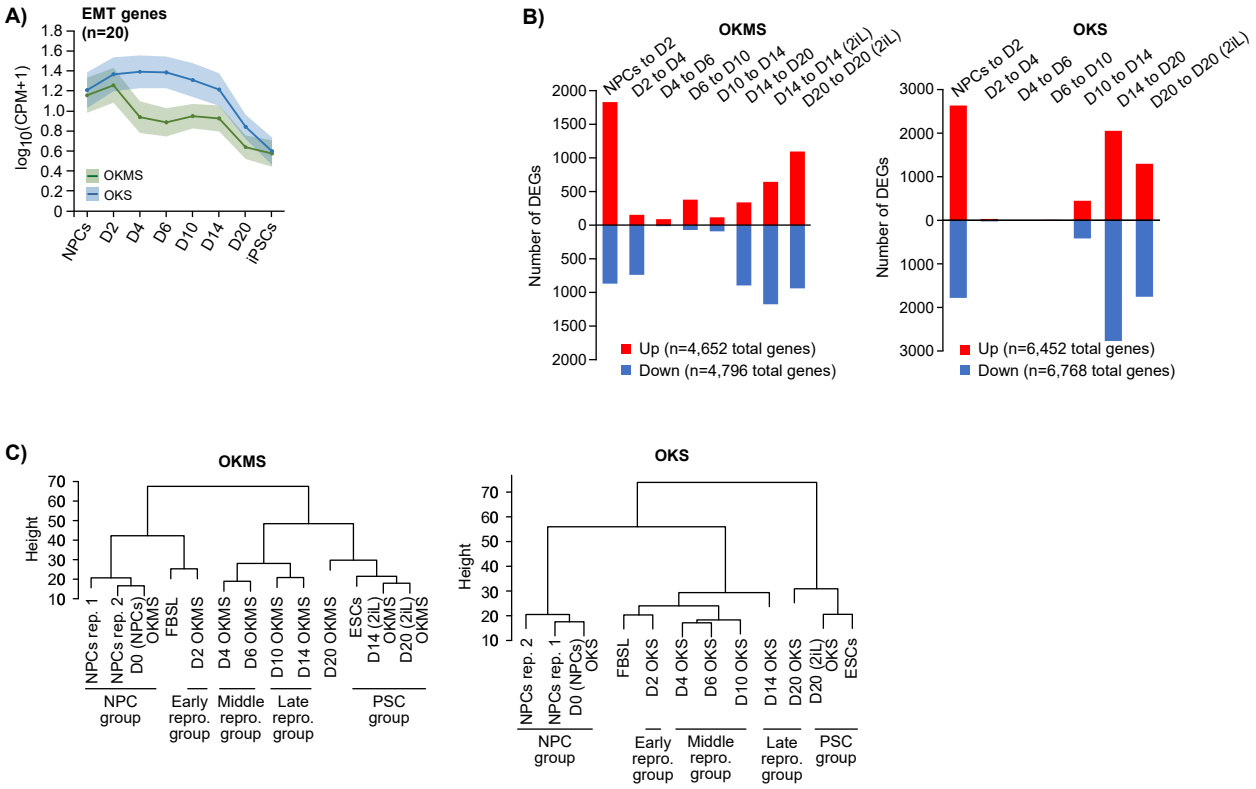
C) Representative image of *Snhg26* lncRNA localization within mouse ESCs in 2iL medium, stained with DAPI (blue staining) and *Snhg26*-specific probes coupled to TSA Vivid dye 650 (red dots). Scalebar = 10 µm.

D) Schematic depicting the well-supported transcripts (blue) of *SNHG26* lncRNA on the latest version of Ensembl database (release 113[84]). kb: kilobases; Mb: Megabases.

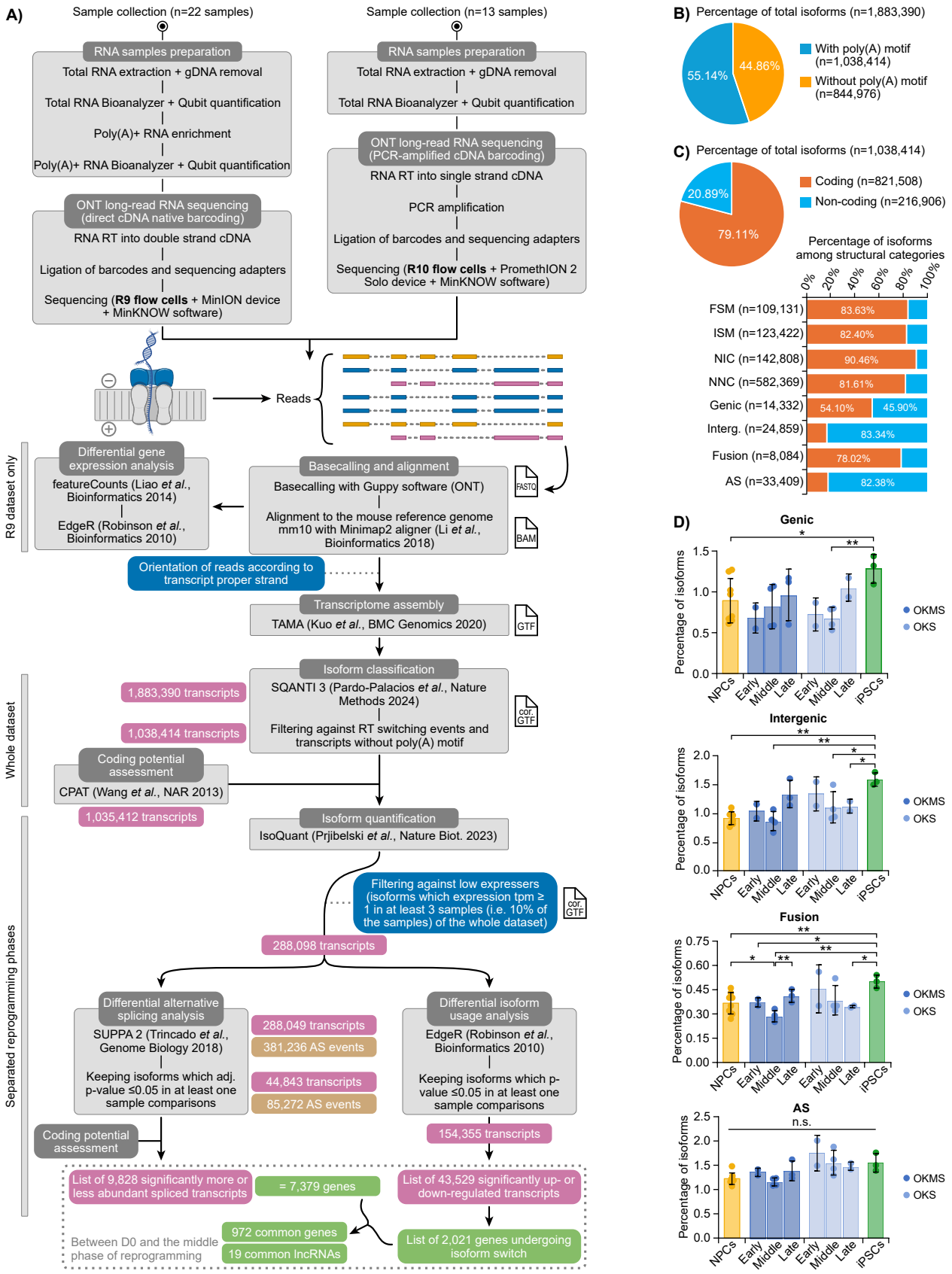
Extended Data Figure 1



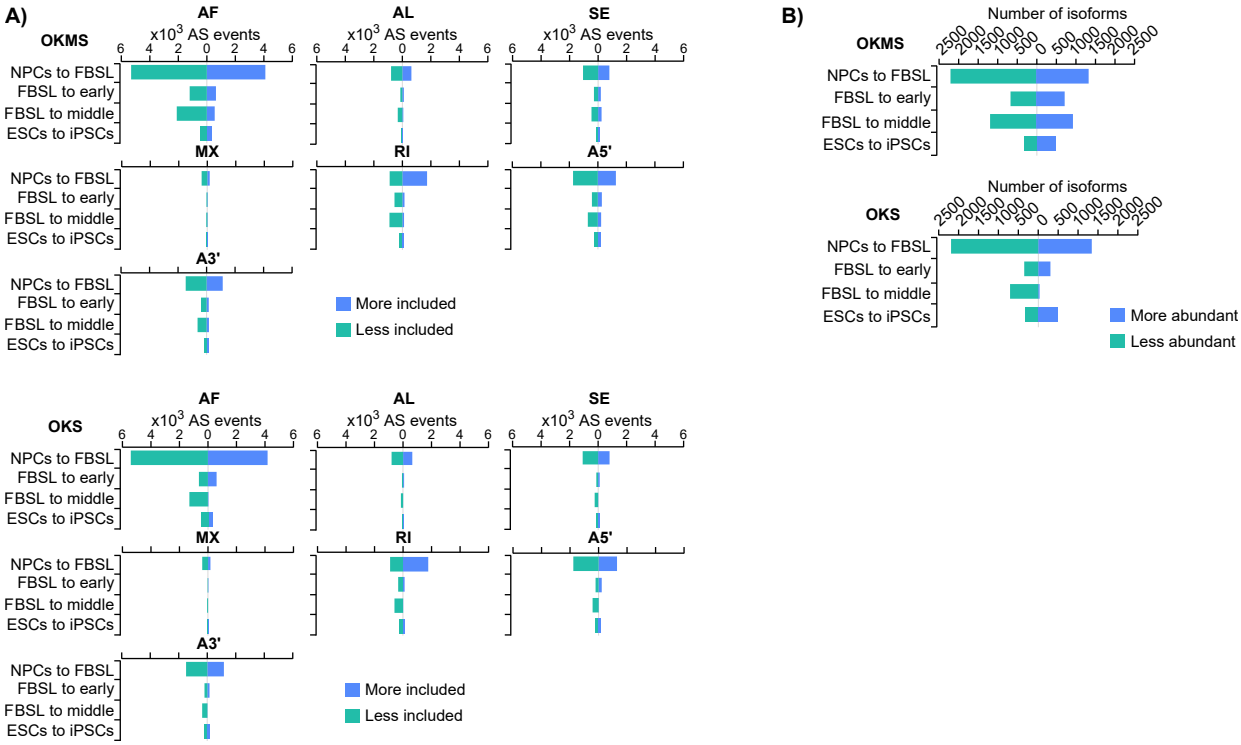
Extended Data Figure 2



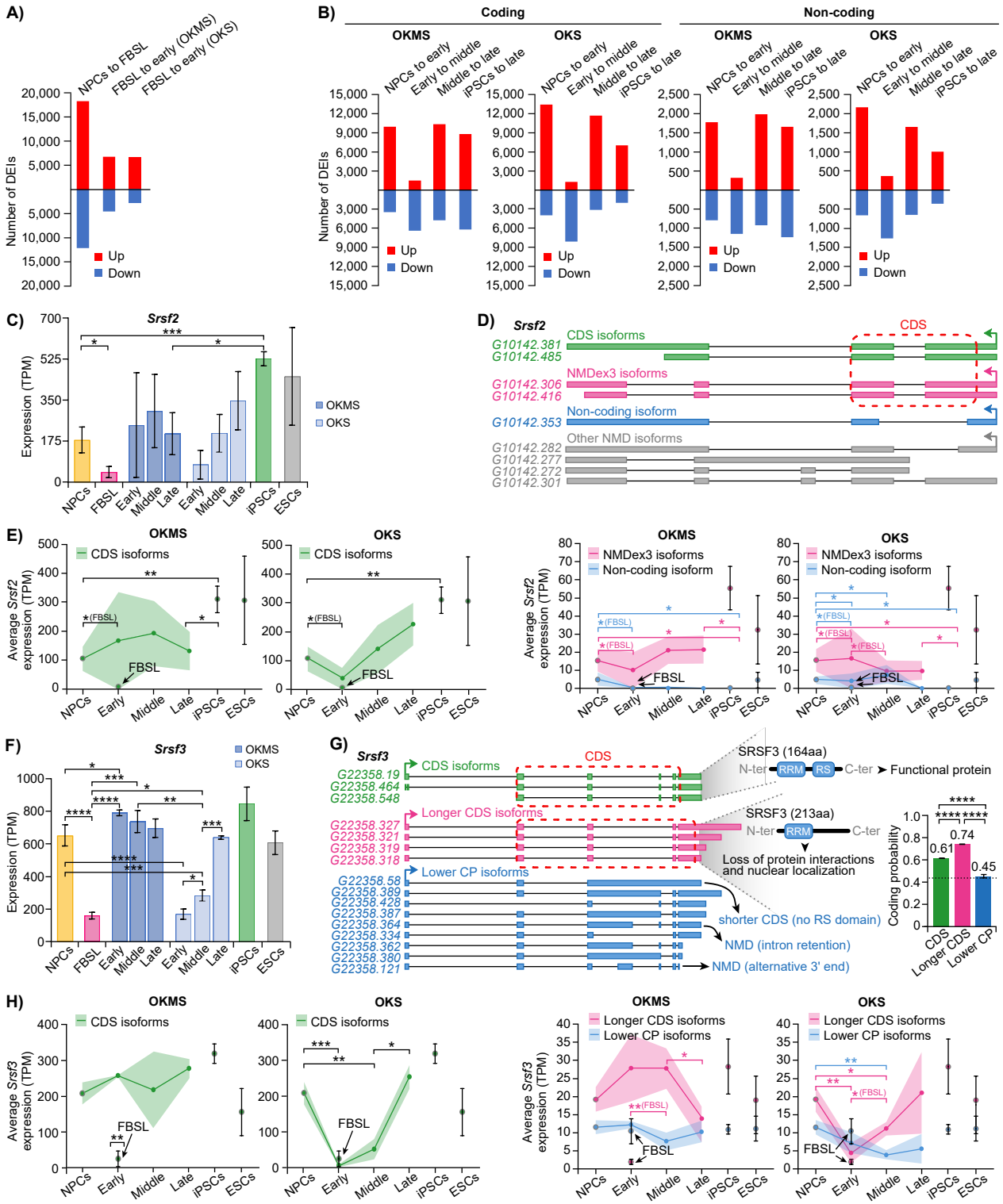
Extended Data Figure 3



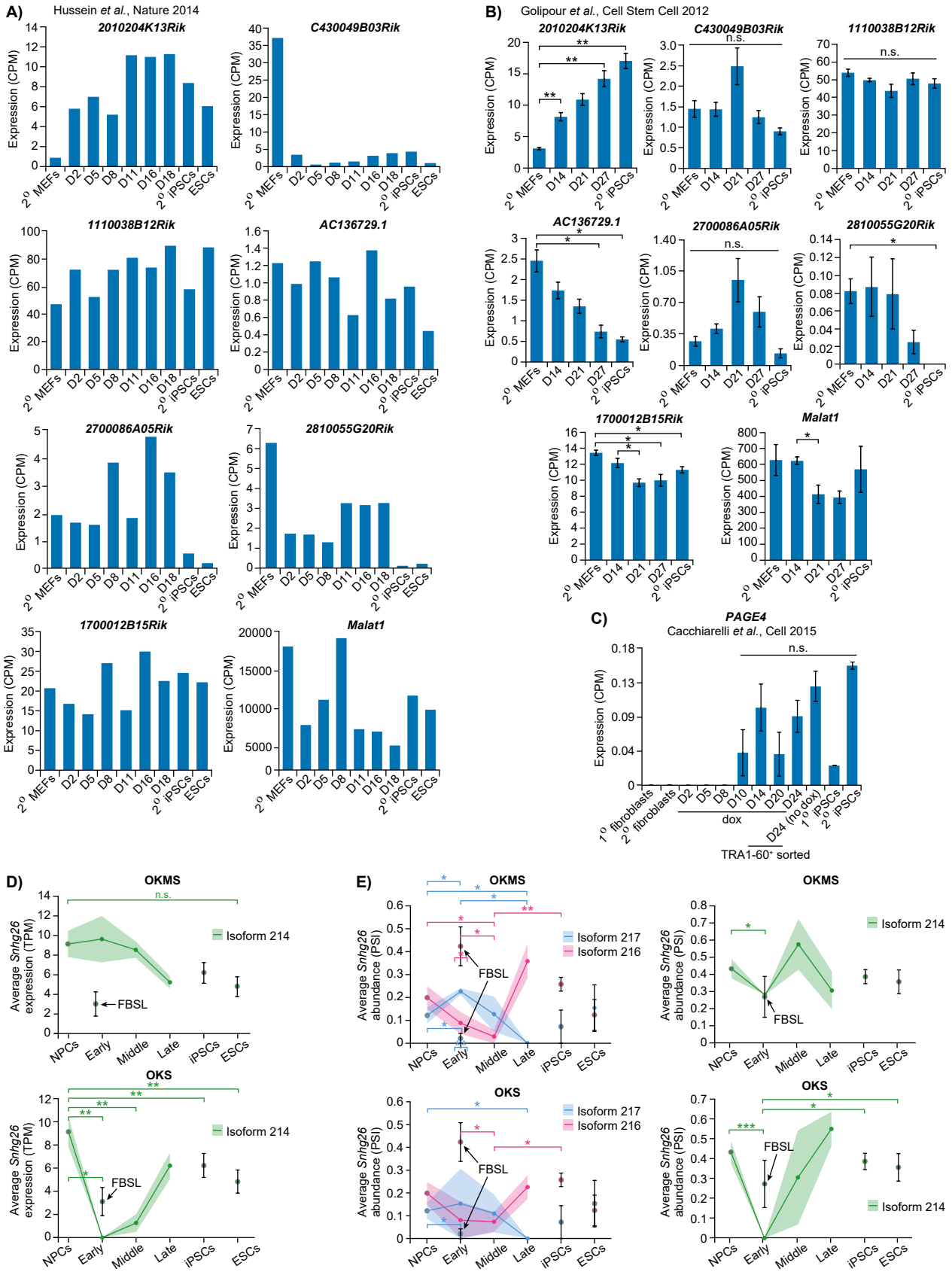
Extended Data Figure 4



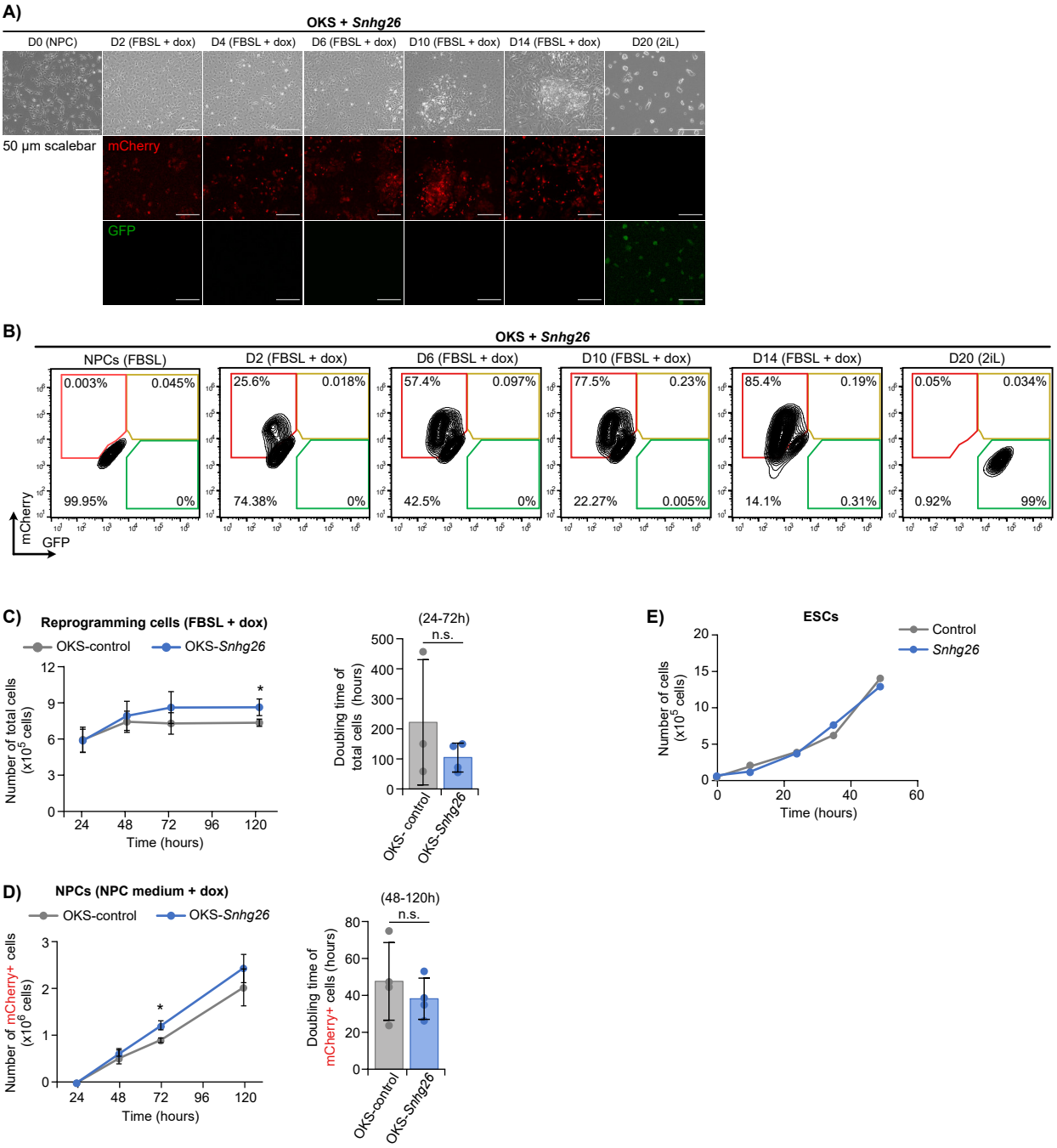
Extended Data Figure 5



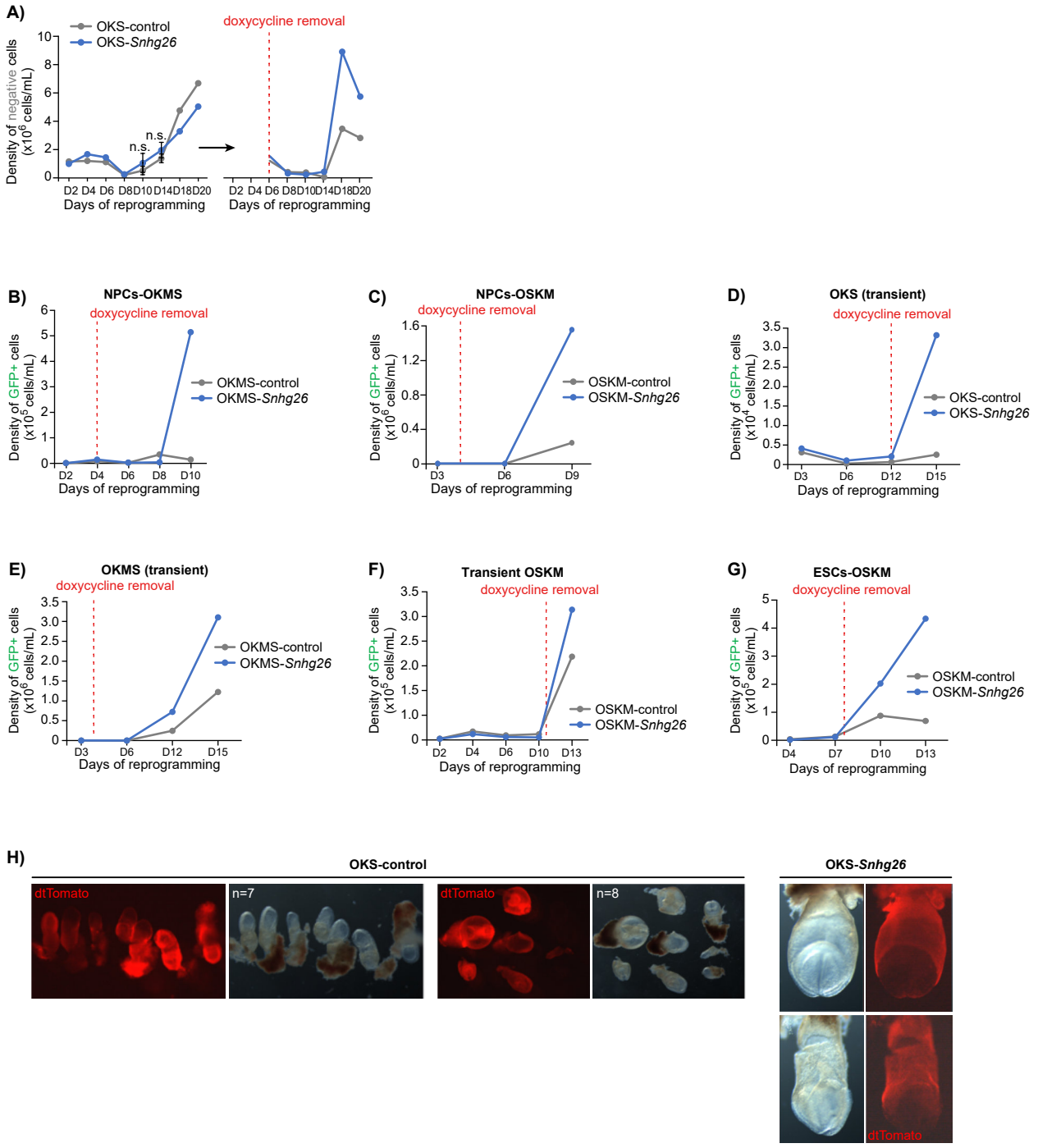
Extended Data Figure 6



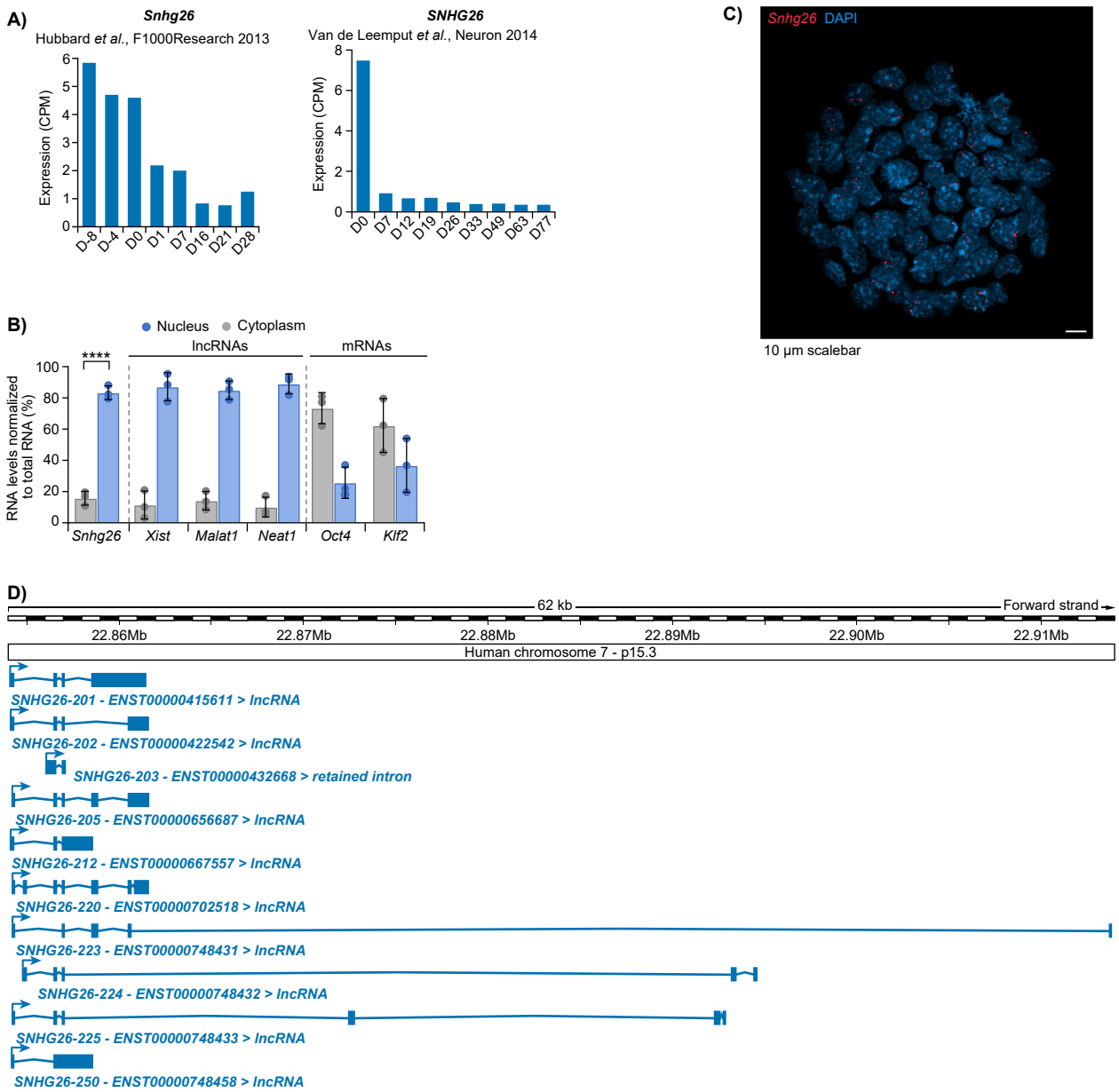
Extended Data Figure 7



Extended Data Figure 8



Extended Data Figure 9



1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

2. 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).

3. Rafelski, S. M. & Theriot, J. A. Establishing a conceptual framework for holistic cell states and state transitions. *Cell* **187**, 2633–2651 (2024).

4. Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

5. Consortium, T. Gte. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

6. He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol.* **21**, 294 (2020).

7. Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nat Biotechnol* 1–13 (2021) doi:10.1038/s41587-021-00936-1.

8. Clough, E. *et al.* NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res.* **52**, D138–D144 (2023).

9. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

10. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535–548 (2018).

11. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

12. Oikonomopoulos, S. *et al.* Methodologies for Transcript Profiling Using Long-Read Technologies. *Front. Genet.* **11**, 606 (2020).

13. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021).

14. 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).

15. Torre, D. *et al.* Isoform-resolved transcriptome of the human preimplantation embryo. *Nat. Commun.* **14**, 6902 (2023).

16. Qiao, Y. *et al.* High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nat. Commun.* **11**, 2653 (2020).

17. Yeo, J.-C. & Ng, H.-H. The transcriptional regulation of pluripotency. *Cell Res* **23**, 20 (2013).

18. Wigger, M. *et al.* Plasticity of the inner cell mass in mouse blastocyst is restricted by the activity of FGF/MAPK pathway. *Sci. Rep.* **7**, 15136 (2017).

19. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

20. Takahashi, K. & Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nat Rev Mol Cell Bio* **17**, nrm.2016.8 (2016).

21. Buganim, Y., Faddah, D. A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14**, 427–439 (2013).

22. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).

23. Hussein, S. M. I. *et al.* Genome-wide characterization of the routes to pluripotency. *Nature* **516**, 198 (2014).

24. Brumbaugh, J., Stefano, B. D. & Hochedlinger, K. Reprogramming: identifying the mechanisms that safeguard cell identity. *Development* **146**, dev182170 (2019).

25. Clancy, J. L. *et al.* Small RNA changes en route to distinct cellular states of induced pluripotency. *Nat. Commun.* **5**, 5522 (2014).

26. Tanaka, Y. *et al.* Transcriptome Signature and Regulation in Human Somatic Cell Reprogramming. *Stem Cell Rep.* **4**, 1125–1139 (2015).

27. Guo, L. *et al.* Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Mol. Cell* **73**, 815-829.e7 (2019).

28. Kim, D. H. *et al.* Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell* **16**, 88–101 (2015).

29. Lujan, E. *et al.* Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521**, 352–356 (2015).

30. Atlasi, Y. *et al.* The translational landscape of ground state pluripotency. *Nat Commun* **11**, 1617 (2020).

31. Aydin, S. *et al.* Genetic dissection of the pluripotent proteome through multi-omics data integration. *Cell Genom.* **3**, 100283 (2023).

32. 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).

33. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442-459.e20 (2017).

34. 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).

35. Panopoulos, A. D. *et al.* The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Res.* **22**, 168–177 (2012).

36. Park, S.-J. *et al.* Metabolome Profiling of Partial and Fully Reprogrammed Induced Pluripotent Stem Cells. *Stem Cells Dev.* **26**, 734–742 (2017).

37. Boyer, L. A. *et al.* Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* **122**, 947–956 (2005).

38. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295 (2011).

39. 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).

40. Subramanyam, D. *et al.* Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 443–448 (2011).

41. Das, S., Jena, S. & Levasseur, D. N. Alternative Splicing Produces Nanog Protein Variants with Different Capacities for Self-renewal and Pluripotency in Embryonic Stem Cells*. *J. Biol. Chem.* **286**, 42690–42703 (2011).

42. Gabut, M. *et al.* An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell* **147**, 132–146 (2011).

43. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241–245 (2013).

44. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1122–1130 (2013).

45. Lu, Y. *et al.* Alternative Splicing of MBD2 Supports Self-Renewal in Human Pluripotent Stem Cells. *Cell Stem Cell* **15**, 92–101 (2014).

46. He, C. *et al.* High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. *Mol Cell* **64**, 416–430 (2016).

47. Dvir, S. *et al.* Uncovering the RNA-binding protein landscape in the pluripotency network of human embryonic stem cells. *Cell Reports* **35**, 109198 (2021).

48. Ohta, S., Nishida, E., Yamanaka, S. & Yamamoto, T. Global Splicing Pattern Reversion during Somatic Cell Reprogramming. *Cell Rep.* **5**, 357–366 (2013).

49. Cieply, B. *et al.* Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. *Cell Rep.* **15**, 247–255 (2016).

50. Hirsch, C. L. *et al.* Myc and SAGA rewire an alternative splicing network during early somatic cell reprogramming. *Genes Dev.* **29**, 803–816 (2015).

51. Toh, C.-X. D. *et al.* RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. *Cell Rep.* **15**, 2597–2607 (2016).

52. Fang, F. *et al.* A distinct isoform of ZNF207 controls self-renewal and pluripotency of human embryonic stem cells. *Nat. Commun.* **9**, 4384 (2018).

53. Kim, Y.-D. *et al.* ESRP1-Induced CD44 v3 Is Important for Controlling Pluripotency in Human Pluripotent Stem Cells. *Stem Cells* **36**, 1525–1534 (2018).

54. Kanitz, A., Syed, A. P., Kaji, K. & Zavolan, M. Conserved regulation of RNA processing in somatic cell reprogramming. *BMC Genom.* **20**, 100 (2019).

55. Vivori, C. *et al.* Dynamics of alternative splicing during somatic cell reprogramming reveals functions for RNA-binding proteins CPSF3, hnRNP UL1, and TIA1. *Genome Biol.* **22**, 171 (2021).

56. Schlaeger, T. M. *et al.* A comparison of non-integrating reprogramming methods. *Nat. Biotechnol.* **33**, 58–63 (2015).

57. Hochedlinger, K. & Plath, K. Epigenetic reprogramming and induced pluripotency. *Development* **136**, 509–523 (2009).

58. 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).

59. David, L. & Hirsch, C. L. How Tets and Cytoskeleton Dynamics MET in Reprogramming. *Cell Stem Cell* **14**, 417–418 (2014).

60. 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).

61. Nakagawa, M. *et al.* Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat. Biotechnol.* **26**, 101–106 (2008).

62. Polo, J. M. *et al.* A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151**, 1617–1632 (2012).

63. 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).

64. Wyman, D. *et al.* A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *Biorxiv* 672931 (2020) doi:10.1101/672931.

65. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**, 396–411 (2018).

66. Pardo-Palacios, F. J. *et al.* SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* 1–5 (2024) doi:10.1038/s41592-024-02229-2.

67. Wong, J. J.-L. & Schmitz, U. Intron retention: importance, challenges, and opportunities. *Trends Genet.* **38**, 789–792 (2022).

68. 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).

69. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Bio* **19**, 143 (2017).

70. Guil, S. *et al.* Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat Struct Mol Biol* **19**, 664–670 (2012).

71. Barman, P., Reddy, D. & Bhaumik, S. R. Mechanisms of Antisense Transcription Initiation with Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis. *Non-Coding RNA* **5**, 11 (2019).

72. Han, H. *et al.* Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. *Mol. Cell* **65**, 539-553.e7 (2017).

73. Venables, J. P. *et al.* MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. *Nat. Commun.* **4**, 2480 (2013).

74. 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).

75. 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 Res.* **51**, 1245–1259 (2023).

76. Lee, J., Park, Y.-J. & Jung, H. Protein Kinases and Their Inhibitors in Pluripotent Stem Cell Fate Regulation. *Stem Cells Int.* **2019**, 1569740 (2019).

77. Brignatz, C. *et al.* Alternative Splicing Modulates Autoinhibition and SH3 Accessibility in the Src Kinase Fyn. *Mol. Cell. Biol.* **29**, 6438–6448 (2009).

78. 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).

79. Utikal, J. *et al.* Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. *Nature* **460**, 1145–1148 (2009).

80. Smith, Z. D., Sindhu, C. & Meissner, A. Molecular features of cellular reprogramming and development. *Nat Rev Mol Cell Bio* **17**, 139–154 (2016).

81. More, D. A. & Kumar, A. SRSF3: Newly discovered functions and roles in human health and diseases. *Eur. J. Cell Biol.* **99**, 151099 (2020).

82. 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).

83. Cacchiarelli, D. *et al.* Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* **162**, 412–424 (2015).

84. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res.* **52**, D891–D899 (2023).

85. Zhong, R. K., Astle, C. M. & Harrison, D. E. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *J. Immunol.* **157**, 138–145 (1996).

86. Ploemacher, R. E., Sluijs, J. P. van der, Voerman, J. S. & Brons, N. H. An in vitro limiting-dilution assay of long-term repopulating hematopoietic stem cells in the mouse. *Blood* **74**, 2755–63 (1989).

87. Sieburg, H. B., Cho, R. H. & Müller-Sieburg, C. E. Limiting dilution analysis for estimating the frequency of hematopoietic stem cells Uncertainty and significance. *Exp. Hematol.* **30**, 1436–1443 (2002).

88. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

89. Koche, R. P. *et al.* Reprogramming Factor Expression Initiates Widespread Targeted Chromatin Remodeling. *Cell Stem Cell* **8**, 96–105 (2011).

90. Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20**, 1131–1139 (2013).

91. 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).

92. 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. *F1000research* **2**, 35 (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 Reports* **26**, 815-824.e4 (2019).

95. Mazid, Md. A. *et al.* Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315–324 (2022).

96. Carlevaro-Fita, J. & Johnson, R. Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. *Mol Cell* **73**, 869–883 (2019).

97. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Bio* 1–17 (2023) doi:10.1038/s41580-022-00566-8.

98. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* **8**, 2180–2196 (2013).

99. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).

100. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113–1117 (2010).

101. Durruthy-Durruthy, J. *et al.* The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat Genet* **48**, 44–52 (2015).

102. Jiapaer, Z. *et al.* LincU Preserves Naïve Pluripotency by Restricting ERK Activity in Embryonic Stem Cells. *Stem Cell Rep* **11**, (2018).

103. Guo, C.-J. *et al.* Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell* (2020) doi:10.1016/j.cell.2020.03.006.

104. 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).

105. Li, D. *et al.* The lncRNA SNHG26 drives the inflammatory-to-proliferative state transition of keratinocyte progenitor cells during wound healing. *Nat. Commun.* **15**, 8637 (2024).

106. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive piggyBac transposase for mammalian applications. *P Natl Acad Sci Usa* **108**, 1531–6 (2011).

107. 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).

108. Viswanathan, S. *et al.* Supplementation-dependent differences in the rates of embryonic stem cell self-renewal, differentiation, and apoptosis. *Biotechnol. Bioeng.* **84**, 505–517 (2003).

109. Mereau, A., Grey, L., Piquet-Pellorce, C. & Heath, J. Characterization of a binding protein for leukemia inhibitory factor localized in extracellular matrix. *J. cell Biol.* **122**, 713–719 (1993).

110. 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**, 2045–2060 (2014).

111. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

112. Coster, W. D. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).

113. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

114. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

115. 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).

116. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **31**, 274–295 (2014).

117. Kuo, R. I. *et al.* Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genom.* **21**, 751 (2020).

118. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74–e74 (2013).

119. Prjibelski, A. D. *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.* **41**, 915–918 (2023).

120. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40 (2018).

121. Heberle, H., Meirelles, G. V., Silva, F. R. da, Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *Bmc Bioinformatics* **16**, 169 (2015).

122. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

123. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2022).