nature genetics

Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance

Alexandre Fort[1,7], Kosuke Hashimoto[1,7], Daisuke Yamada[2], Md Salimullah[1], Chaman A Keya[1], Alka Saxena[1,6], Alessandro Bonetti[1], Irina Voineagu[1,6], Nicolas Bertin[1,6], Anton Kratz[1], Yukihiko Noro[1], Chee-Hong Wong[3], Michiel de Hoon[1], Robin Andersson[4], Albin Sandelin[4], Harukazu Suzuki[1], Chia-Lin Wei[3], Haruhiko Koseki[2], The FANTOM Consortium[5], Yuki Hasegawa[1], Alistair R R Forrest[1] & Piero Carninci[1]

The importance of microRNAs and long noncoding RNAs in the regulation of pluripotency has been documented; however, the noncoding components of stem cell gene networks remain largely unknown. Here we investigate the role of noncoding RNAs in the pluripotent state, with particular emphasis on nuclear and retrotransposon-derived transcripts. We have performed deep profiling of the nuclear and cytoplasmic transcriptomes of human and mouse stem cells, identifying a class of previously undetected stem cell–specific transcripts. We show that long terminal repeat (LTR)-derived transcripts contribute extensively to the complexity of the stem cell nuclear transcriptome. Some LTR-derived transcripts are associated with enhancer regions and are likely to be involved in the maintenance of pluripotency.

Pervasive transcription of mammalian genomes into various long, short, protein-coding and noncoding transcript classes is now an accepted, fundamental observation made by several large-scale studies[1–5]. These projects emphasize the need for complementary high-throughput technologies coupled with integrative bioinformatics approaches to characterize this large diversity of RNA species. These reports focused mostly on the polyadenylated fraction of total RNA, which is dominated by cytosolic RNAs. Recent studies have analyzed the transcriptomes of different subcellular compartments, showing that the nucleus hosts a vast collection of intergenic and antisense transcripts[2,6]. Adding to this complexity, we have previously shown that up to 30% of human and mouse transcription start sites (TSSs) are located in transposable elements and that they exhibit clear tissue-specific and developmental stage–restricted expression patterns[7]. Despite this identification of a diverse abundance of transcripts derived from noncoding and repetitive elements, relatively little is known of the functions of these RNAs.

Functional roles for human repeat-derived transcripts have been investigated in humans, particularly in early-stage embryos and embryonic stem cells (ESCs), where expression of HERV-H has been described as a marker of pluripotency[8], whereas, in mice, MuERV-L elements have been shown to trigger early embryonic development[9]. More recently, MuERV-L elements have been shown to fine tune the genomic network of totipotent cells at the two-cell stage[10]. In addition, LTR-associated binding sites for stem cell–specific transcription factors[11,12] and LTR sequences enriched in stem cell–specific long noncoding RNAs (lncRNAs)[13] have been reported. Taken together, these findings point to a likely role for repeat-associated noncoding transcripts in the maintenance of pluripotency and lineage commitment.

ESCs and induced pluripotent stem cells (iPSCs) are increasingly being used for drug screening as well as for cell-based models of numerous pathologies and regenerative medicine. For such cellular models, understanding the mechanisms that maintain stemness and promote differentiation is critical. Genes implicated in the maintenance of pluripotency have been identified by multiple studies, forming a well-documented regulatory network[14]. Long intergenic noncoding RNAs[15–17] (lincRNAs) and microRNAs[18–20] (miRNAs) have also been shown to be part of this complex regulatory system. However, the role of noncoding RNAs and retrotransposon-derived transcripts in the genetic network regulating pluripotency status is likely to be underestimated, as previous studies have focused only on known noncoding transcripts[15,16]. It is therefore necessary to establish a comprehensive noncoding transcriptome from a representative collection of human and mouse stem cells, which can be used as a reference for comparison to and between newly engineered cell models.

[1]Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan. [2]Laboratory for Developmental Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [3]Sequencing Technology Group, Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California, USA. [4]Bioinformatics Centre, Department of Biology and Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark. [5]A full list of members and affiliations appear in the **Supplementary Note**. [6]Present addresses: National Institute for Health Research, Biomedical Research Centre, Genomics Core Facility, Guy's Hospital, London, UK (A. Saxena), School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia (I.V.), and Cancer Science Institute of Singapore, National University of Singapore, Singapore (N.B.). [7]These authors contributed equally to this work. Correspondence should be addressed to P.C. (carninci@riken.jp).

Received 21 June 2013; accepted 2 April 2014; published online 28 April 2014; doi:10.1038/ng.2965

To this end, we performed deep profiling of nucleus-enriched and cytoplasmic RNA fractions from a representative set of human and mouse stem cell lines using four complementary high-throughput sequencing technologies. Consequently, we could identify and characterize several thousand antisense, intergenic and intronic transcripts, including stem state–specific repeat-associated RNAs, principally localized to the nucleus. Strikingly, a large fraction of these newly identified transcripts originate in LTR elements, and we show for four candidates that their expression is associated with the maintenance of pluripotency. Additionally, LTR-derived transcripts are often associated with distal regulatory elements.

RESULTS

Characterization of non-annotated stem transcripts

To generate a comprehensive and representative catalog of the transcripts expressed in mammalian pluripotent stem cells, we selected 11 different mouse and human pluripotent cell lines, including ESCs and iPSCs (**Table 1**). In detail, ESC lines derived from three different inbred mouse strains as well as three human ESC lines were used in this study. In addition, iPSCs reprogrammed from human fibroblasts (**Supplementary Fig. 1**), human B lymphocytes (D.Y. and H.K., unpublished data), mouse B lymphocytes (D.Y. and H.K., unpublished data), mouse T lymphocytes (**Supplementary Fig. 1**) and mouse fibroblasts[21] were included as additional pluripotent lines. The differentiated cell types used for iPSC derivation were integrated into the analysis as controls. With the aim of detecting abundant as well as rare compartment-specific transcripts, we analyzed nucleus-enriched and cytoplasmic RNA fractions from all 11 pluripotent lines and the 6 differentiated cell types. Deep transcriptome profiling of these 34 samples was conducted, including analysis of genome-wide TSS activities using CAGE (cap analysis of gene expression[22]), transcript assemblies based on data from CAGEscan (nanoCAGE combined with paired-end sequencing[23]) and RNA sequencing (RNA-seq)[24]. Lastly, post-transcriptional processing events were assessed using data from short RNA-seq (**Table 1**). We created an extensive transcriptome profile on the basis of 1.38 and 1.27 billion tag sequences and 102,495 and 106,027 CAGE tag clusters (TSSs; **Supplementary Data Sets 1** and **2**) as well as 306,358 and 562,430 assembled transcripts for mice and humans, respectively. The criteria used to create a strict set of TSSs included considering only CAGE tag clusters uniquely mapping to the reference genomes that were detected in two samples or more and were expressed above one tag per million (tpm) in at least one sample. Notably, the vast majority of CAGE clusters were expressed in multiple stem cell samples, with half of them found in all ESC and iPSC lineages (**Supplementary Fig. 2a**,**b**). Corroborating previous reports[2,4,6], high nuclear transcriptome complexity was observed for both species, with 46–55% of detected CAGE clusters found in nucleus-enriched samples only, whereas transcripts observed only in the cytoplasm counted for 5.2–7.5% of transcripts and the remaining transcripts (39.8–46.5%) were observed in libraries from both cellular compartments (**Fig. 1a**). This distribution of transcripts was quantified by estimating the number of CAGE clusters expressed in each sample after correcting for sequencing depth (**Fig. 1b**). Consistent with previous studies[14,25], we found that stem cells expressed a greater diversity of transcripts (30% more complexity on average) than differentiated cells. In addition, hierarchical clustering on the basis of the expression patterns for CAGE clusters separated differentiated and stem cell samples, as expected, with these groups forming subclusters corresponding to nuclear and cytoplasmic fractions (**Supplementary Fig. 2c**,**d**). Cufflinks[26] transcript assemblies based on RNA-seq data, performed for four ESC lines, also showed greater nuclear complexity, with 77.1% and 55.2% of mouse and human transcripts, respectively, identified only in the nucleus (**Supplementary Fig. 3**).

To identify new RNAs implicated in transcriptional regulation specifically in stem cells, we compared transcript expression in ESC and iPSC lineages with that in three differentiated cell types (fibroblasts, B lymphocytes and T lymphocytes) for nuclear and cytoplasmic transcript sets separately. We identified a total of 15,059 (out of 102,495; 14.7%) and 8,254 (out of 106,027; 7.8%) CAGE clusters expressed at significantly higher levels (false discovery rate (FDR) < 0.01, calculated with edgeR[27]) in mouse and human stem cells, respectively (**Supplementary Fig. 4a–d**). These CAGE clusters were classified as 'nuclear' (mouse, n = 8,601; human, n = 2,804) or 'cytoplasmic' (mouse, n = 1,915; human, n = 1,544) when identified

Table 1 Cell lines used for deep transcriptome profiling and sequencing depth

			Aligned tags (×10^6)			
Cell line (clone name)	Cell type	Strain or sex	CAGE (N/C)	CAGEscan (N/C)	Short RNA-seq (N/C)	RNA-seq (N/C)
Mouse						
mESR08 (Nanog^(βgeo/+)ES)	ESC	129 SV Jae	19.7/16.6	20.5/27.3	26.9/12.3	50.0/46.3
mESB6G-2	ESC	C57BL/6	16.2/16.2	23.1/7.3	38.1/20.4	77.4/60.9
mESFVB-1	ESC	FVB	19.9/16.2	19.8/11	31.7/18.1	
miPS.F (iPS_MEF-Ng-20D17)	iPSC	C57BL/6	17.9/14.8	20.2/28.1	25.5/22.5	
miPS.B (iPS_LymB_44.1B4e)	iPSC	C57BL/6	14.7/16.7	21/23.1	26.9/14.0	
miPS.T (iPS_LymT_i103 H12)	iPSC	C57BL/6	15.1/17.4	8.9/25.5	28.1/22.1	
MEF (MEF_Ng-20D17)	Fibroblast	C57BL/6	23.9/15.8	29.5/20.3	22.2/25.3	
Primary B lymphocytes	B cell	C57BL/6	18.0/16.2	15.8/27.1	28.0/21.1	
Primary T lymphocytes	T cell	C57BL/6	18.4/15.5	27.9/25.2	26.4/21.1	
Human						
KhES-1	ESC	Female	22.9/16.2	24.3/27.2	21.6/20.0	
KhES-2	ESC	Female	19.4/15.4	29.4/33.1	23.3/19.5	105/46.4
KhES-3	ESC	Male	20.0/16.3	41.6/17.0	14.6/15.6	49.3/33.3
hiPS.F (iPS_HDF-f_hi6)	iPSC	Male	19.6/15.1	26.6/8.6	28.1/19.3	
hiPS.B (iPS_LymB_hi68)	iPSC	Male	19.2/18.6	28.3/19.1	26.4/17.7	
HDF-f	Fibroblast	Male	10.7/26.6	26.5/3.5	18.1/28.6	
Primary B lymphocytes	B cell	Male	16.3/8.4	2.0/2.8	15.6/14.1	
Primary T lymphocytes	T cell	Male	19.5/27.5	26.2/25.9	61.3/54.0	

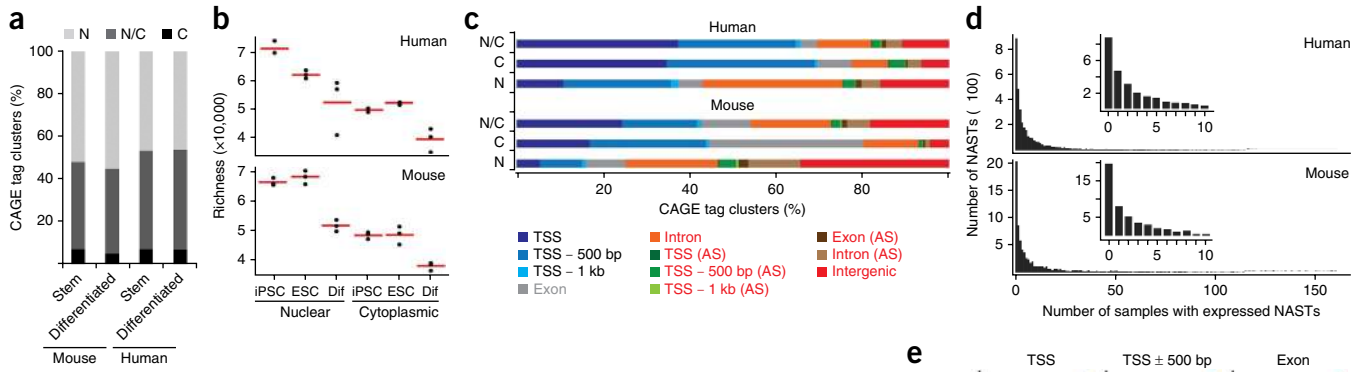

Figure 1 High nuclear complexity of the stem cell transcriptome and identification of stem cell–specific transcripts. (**a**) Proportions of the CAGE tag clusters expressed in stem or differentiated cells identified in the nuclear (N) or cytoplasmic (C) compartments or in both (N/C). (**b**) Comparison of transcriptional complexity in the nuclear and cytoplasmic compartments of stem (ESC and iPSC) and differentiated (Dif) cells, calculating the number of CAGE clusters expressed in each sample after correcting for sequencing depth (richness). Red bars represent mean values. (**c**) Annotation of CAGE tag clusters significantly upregulated in stem cells (edgeR[27], FDR < 0.01). NAST candidates are represented in red text. AS, antisense. (**d**) Number of tissues and differentiated cell type samples from the FANTOM5 expression atlas[29] in which NASTs are expressed. Insets, enlarged plots of the first ten bins. Bin width = 1 sample. (**e**) Nuclear (*x* axis) and cytoplasmic (*y* axis) normalized expression for the mouse CAGE clusters significantly overexpressed in stem cells.

as significantly overexpressed in stem cells in only 1 of the 2 compartments and were classified as 'common' (mouse, n = 4,543, human, n = 3,906) if found to be significantly upregulated in both compartments (**Supplementary Fig. 4e**). These stem cell–specific CAGE clusters were annotated using all available annotation sources (Online Methods). A total of 8,873 (out of 15,059; 58.9%) mouse and 3,042 (out of 8,254; 36.9%) human stem cell–specific CAGE clusters were found either in the antisense direction relative to annotated genes or residing in intronic and intergenic regions (**Fig. 1c**). We consider this fraction of stem cell–specific TSSs not directly associated with known genes to represent potentially new stem cell–specific RNAs and name them 'non-annotated stem transcripts' (NASTs; **Supplementary Data Sets 1** and **2**). We compared our set of NASTs with two recent catalogs of human[13] and mouse[28] lncRNAs and observed that 3.8% and 12% of human and mouse NASTs, respectively, were located within a 500-bp window centered on the TSS of an already reported lncRNA. The frequency of overlap increased to 7.9% and 19.9% when we compared human and mouse NAST TSSs to all reported lncRNA exons, keeping a permissive 500-bp window, indicating that current lncRNA gene models are potentially incomplete.

Notably, the majority of NASTs were found to be expressed in all stem cell lines used in this study and did not show cell line–specific expression patterns (**Supplementary Fig. 5a–d**). In addition, 60–85% of the NAST TSS positions were independently confirmed by CAGEscan 5′ tags (**Supplementary Fig. 5e**). We confirmed the stem cell specificity of NASTs by examining their expression across sets of 165 human and 120 mouse samples selected from the FANTOM5 expression atlas[29] (covering adult, fetal and embryonic tissues as well as differentiated primary cell types). These NASTs appeared to be expressed in only a few differentiated cell types, notably mainly in testis (**Fig. 1d**). These expression patterns contrast with the ones observed for the annotated CAGE clusters found to be significantly overexpressed in stem cells that were expressed in multiple FANTOM5 samples (**Supplementary Fig. 5f**). In addition, NASTs were clearly more abundant in the nucleus, as shown by their higher nuclear expression levels compared to cytoplasmic levels (**Fig. 1e** and **Supplementary Fig. 5g–i**). Finally, NAST identification was also supported by signal enrichments in global nuclear run-on sequencing (GRO-seq; indicating the levels of transcriptionally engaged RNA polymerase II across the genome) at NAST promoters, when using data for human[30] and mouse[31] ESCs (**Supplementary Fig. 5j,k**).

Taken together, our results suggest that the transcriptional complexity of stem cells has historically been underestimated. Our analyses indeed identify a large proportion of TSSs, specifically expressed in stem cells, that are enriched in the nucleus and are not associated with any previously annotated transcripts.

Further characterization of NASTs showed that these RNAs are mainly species specific and overlap repeated regions. Indeed, we could detect syntenic conservation for only 43% of human and 40% of mouse NASTs using available whole-genome alignment tools (liftOver[32], requiring 80% sequence homology), whereas 74% and 83% of known (already annotated) human and mouse stem cell–specific CAGE clusters, respectively, could be aligned to both species (**Fig. 2a**). The non-conserved fraction of NASTs was strikingly enriched for repetitive elements, which were often lineage specific and thus imply a lack of syntenic conservation. These features are in agreement with a report that each mammalian clade has evolved its own distinct repertoire of lincRNAs[16].

We next used Encyclopedia of DNA Elements (ENCODE) chromatin immunoprecipitation and sequencing (ChIP-seq) data[5] for histone marks in ESCs (human, H1-hESC; mouse, ES-Bruce4 and ES-E14) to classify the genomic loci of NASTs as promoters, enhancers, repressed or overlapping regions carrying histone marks specific to introns and exons (Online Methods). We found that 80% of human and mouse NASTs carried histone marks for enhancers or promoters (**Fig. 2b** and **Supplementary Fig. 6**), thus providing independent confirmation of the transcriptionally active state of NAST loci in stem cells.

Comparing NASTs to the TSSs of known stem cell–specific genes, we observed that they tended to be weakly expressed, were shorter and were less likely to be processed into short RNAs. Both mouse and human NASTs were generally expressed at lower levels than transcripts from annotated TSSs (**Fig. 2c**), with a stronger trend

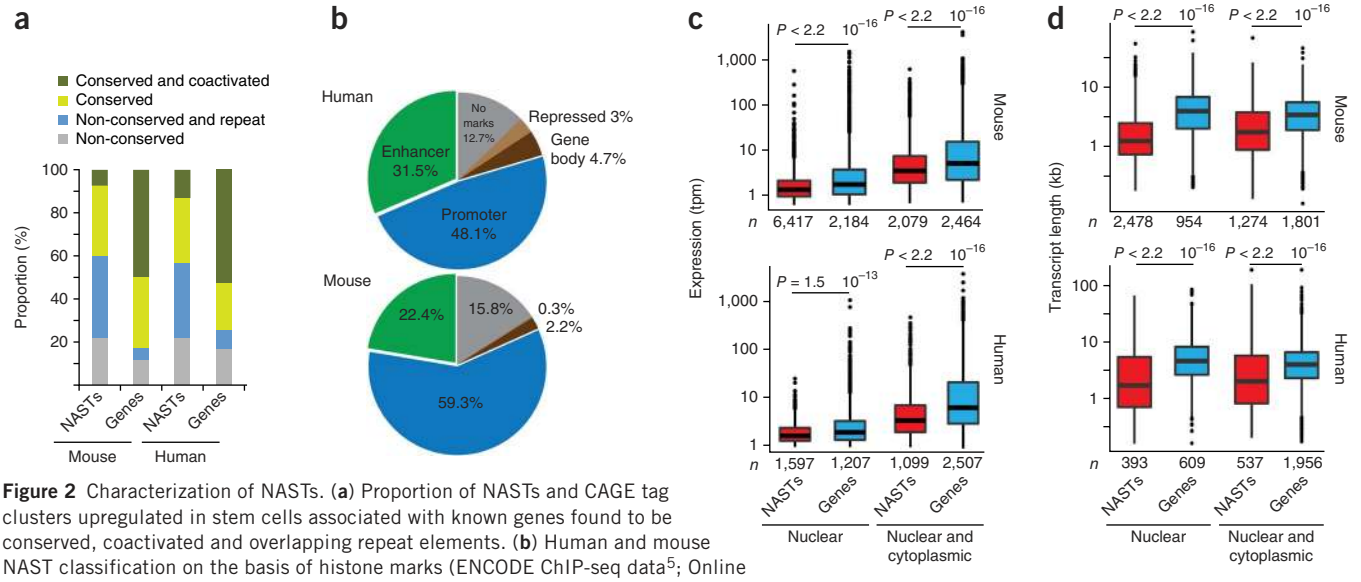

Figure 2 Characterization of NASTs. (**a**) Proportion of NASTs and CAGE tag clusters upregulated in stem cells associated with known genes found to be conserved, coactivated and overlapping repeat elements. (**b**) Human and mouse NAST classification on the basis of histone marks (ENCODE ChIP-seq data[5]; Online Methods). (**c**,**d**) Normalized expression (**c**) and transcript length (**d**) derived from RNA-seq data for NASTs and CAGE tag clusters upregulated in stem cells associated with known genes. Boxes indicate 25th and 75th percentiles, bold bars indicate medians and, whiskers represent 5th and 95th percentiles. *P* values from Wilcoxon and Mann-Whitney two-sided tests are shown, and *n* is the number of CAGE clusters or transcripts per group.

observed for NASTs carrying promoter-associated histone marks (**Supplementary Fig. 7a**,**b**). We confirmed experimentally, using quantitative RT-PCR (qRT-PCR), that weakly expressed NASTs (close to 1 tpm) were present at more than one copy per cell (**Supplementary Fig. 7c**). On the basis of transcripts assembled from RNA-seq data (**Supplementary Data Sets 1** and **2**), NASTs were significantly (Wilcoxon and Mann-Whitney two-sided test, $P < 2.2 \times 10^{-16}$) shorter than the transcripts associated with annotated TSSs (**Fig. 2d**), also when considering their relative expression levels (**Supplementary Fig. 7d**). Finally, only 9–33% and 27–53% of mouse and human NAST genomic loci, respectively, as defined by CAGEscan data, were found to overlap with clusters of short RNAs (**Supplementary Data Sets 1** and **2**), whereas 32–64% of mouse and 42–72% of human transcripts associated with annotated TSSs had evidence of processing into short RNAs (**Supplementary Fig. 7e**).

In summary, our genome-wide survey of promoter activity, transcript assemblies from RNA-seq data and short RNA-seq results clearly show that a large fraction of the stem cell transcriptome is composed of not-yet-annotated transcripts (or NASTs for non-annotated stem transcripts), residing mainly in the nucleus. NASTs are less conserved and overlap more frequently with repetitive elements. Furthermore, NASTs are shorter and are expressed at lower levels than known RNAs. Finally, the vast majority of NASTs are supported by active promoter and enhancer histone marks.

Stem cell–specific LTR-derived transcripts

Our analysis showed that NAST promoters were located more often than expected by chance (Fisher's exact test, Bonferroni corrected, $P < 0.05$) in specific LTR retrotransposon families (**Supplementary Table 1**). Notably, these associations were not observed for non-annotated (antisense, intronic and intergenic) transcripts specific to differentiated cells. In addition, LTR-associated promoters, selected on the basis of histone marks, were expressed in stem cells at significantly higher levels among NASTs than promoters not associated with repeats (Wilcoxon and Mann-Whitney two-sided test, $P \leq 0.0036$); this was not observed for annotated genes (**Supplementary Fig. 8a**,**b**). Our observations on NAST promoters suggest a generalization of the observation made by Kelley and Rinn[13] for 9,241 human and 981 mouse lncRNAs that were found to be strongly associated with LTR elements. Indeed, NASTs were strongly associated with the ERVK and MaLR LTR subfamilies in mice and with ERV1 in humans (**Fig. 3a**). In the vast majority of cases, the LTR elements identified in our analyses were so-called solitary elements, distinct from full-length elements carrying viral ORFs, such as mouse intracisternal A particles (IAPs) and human HERV-K[33,34].

A recent report indicated that the histone methyltransferase SETDB1 mediates repression of numerous noncoding and repetitive elements in mouse ESCs, regulating trimethylation of histone H3 at lysine 9 (H3K9me3)[35]. In light of this observation, we examined the presence of this histone mark at NAST loci, finding that NASTs associated with mouse ERVK, mouse MaLR and human ERV1 elements were deprived of H3K9me3 marks, whereas non-expressed elements were indeed carrying these repressive marks (ENCODE ChIP-seq[5] data for H1-hESC and ES-Bruce4 cells; **Fig. 3b**,**c** and **Supplementary Fig. 8c**). These findings suggest specific transcriptional regulation allowing LTR-associated NASTs to escape the global repeat repression pathway shown to be active in ESCs. In addition, enrichment for stem cell transcription factors (NANOG, SOX2 and OCT4 (also known as POU5F1)) bound at NAST loci associated with mouse ERVK, mouse MaLR and human ERV1 elements appeared greater than for the non-expressed elements (ENCODE[5] ChIP-seq data for H1-hESC and ES-Bruce4 cells; **Supplementary Fig. 8d**–**f**).

To confirm the expression of retrotransposon-derived transcripts in stem cells, we performed differential expression analyses focusing exclusively on repetitive elements, including CAGE tags that mapped to multiple genomic loci, which were previously excluded. For this purpose, expression values for each repeat family and subfamily were assessed by mapping CAGE tags to all repeat elements in the human and mouse genomes, as defined by RepeatMasker[36]. When considering expression values calculated for nuclear samples, the ERVK and MaLR families appeared to be significantly more highly expressed in mouse stem cells ($P = 0.0057$, two-sided *t* test) (**Fig. 3d**). In human cells, ERV1 and ERVK elements showed similar trends, being expressed at higher levels in stem cells than in differentiated cells (**Supplementary Fig. 9a**), as expected in light of recent mouse ESC transcriptome data[35,37]. These results were further confirmed by analyzing only

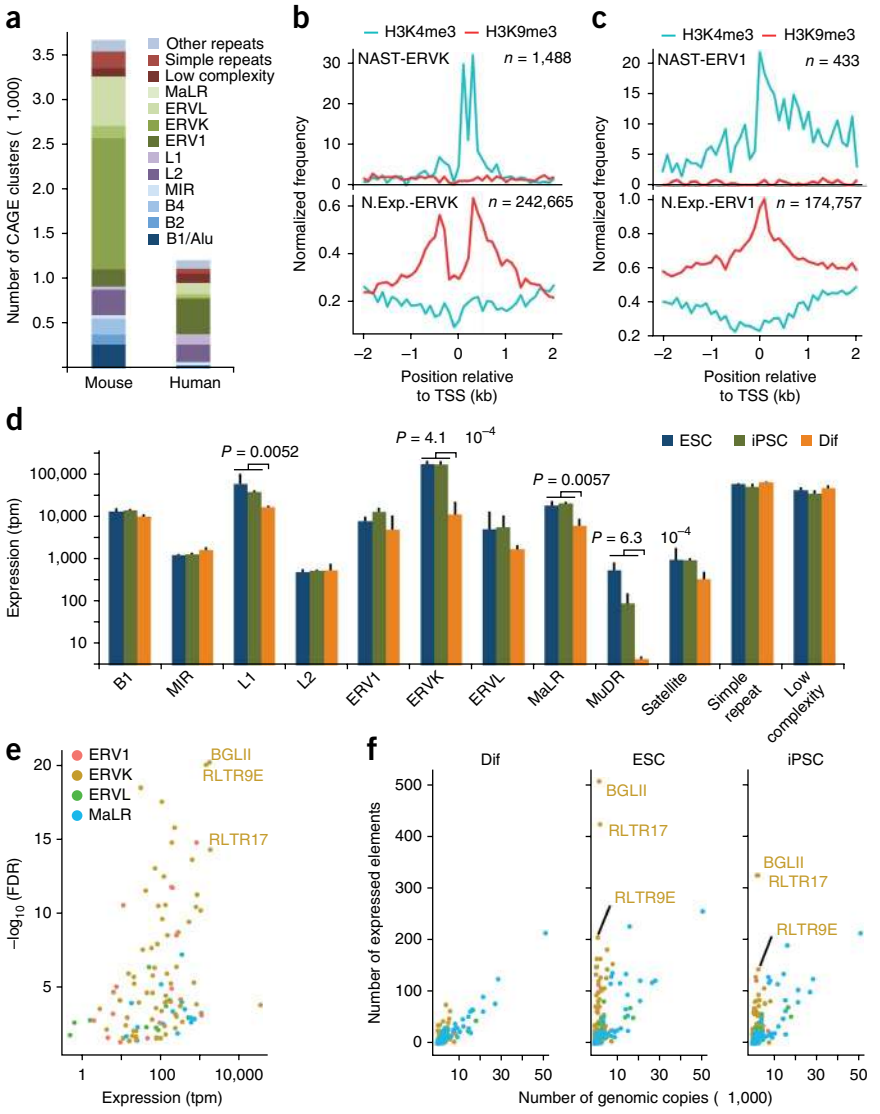

Figure 3 LTR-derived transcripts enriched in NASTs. (**a**) Repeat composition of NASTs. (**b**,**c**) Frequency plots of normalized ChIP-seq (ENCODE data[5]) tag counts for H3K4me3 (promoter) and H3K9me3 (repressed) marks at NAST loci associated with mouse ERVK (**b**) and human ERV1 (**c**) elements. N.Exp., not expressed. (**d**) Normalized expression values for repeat families in mouse ESCs, iPSCs and differentiated cells. Error bars, s.d. Indicated *P* values are from two-sided, Bonferroni corrected *t* tests; ESCs and iPSCs, $n = 6$; differentiated cells, $n = 3$. (**e**) Normalized expression for selected mouse subfamily repeats is plotted against associated FDR (calculated with edgeR[27]). (**f**) The numbers of repeat elements corresponding to at least five CAGE tags for mouse LTRs are plotted against their copy numbers in the genome.

CAGE clusters carrying promoter-associated histone marks (as defined in **Fig. 2b**) and overlapping mouse ERVK or MaLR and human ERV1 elements (**Supplementary Fig. 9b–d**). We then focused on the subfamily level and identified the ERVK elements BGLII, RLTR9E and RLTR17 as the most significantly enriched (FDR calculated by edgeR[27]: BGLII, 7.94×10^{-21}; RLTR9E, 1.67×10^{-20}; RLTR17, 5.80×10^{-15}) in stem cells that showed the highest expression levels (**Fig. 3e**). In human stem cells, LTR7 elements carrying the promoter for the downstream full-length HERVH-int element, as well as LTR7B and LTR7Y elements, were clearly expressed at the highest levels and were the most statistically significant (FDR calculated by edgeR[27]: LTR7, 1.88×10^{-28}; LTR7B, 1.29×10^{-28}; LTR7Y, 4.80×10^{-21}), corroborating recent reports[8,13] (**Supplementary Fig. 9e**). Notably, these highly expressed mouse ERVK and human ERV1 elements were not among the most abundant in the corresponding genome in terms of copy number (**Fig. 3f** and **Supplementary Fig. 9f**), suggesting that we are observing regulated transcription events and not products of random pervasive transcription.

In summary, these analyses suggest that a large fraction of NAST promoters are associated with a few specific subfamilies of mouse ERVK and human ERV1 elements.

Stem cell–specific LTR transcripts associate with enhancers

We next examined the genomic distribution of the tags originating from stem cell–specific LTR-associated promoters for the most highly expressed mouse ERVK and human ERV1 subfamilies. Interestingly, we observed a bidirectional pattern previously reported as a landmark of enhancer regions[38]. Indeed, specifically in ESCs and iPSCs, mouse BGLII elements carried a sharp sense promoter in the center and an antisense (relative to the repeat element orientation) promoter at their 5′ end (**Fig. 4a**). A similar pattern was observed for RLTR17 elements, although promoters appeared broader when compared with those of BGLII elements (**Fig. 4a**). In contrast, for RLTR9E elements, we found a broad distribution of stem cell–specific sense-orientation CAGE tags consisting of multiple promoters (**Supplementary Fig. 10a**). In human stem cells, distinct stem cell–specific promoters were identified in ERV1 elements. A sense-orientation promoter, located close to the 3′ end of LTR7 elements, was also observed upstream of full-length HERVH-int elements (**Supplementary Fig. 10b**).

As part of FANTOM5, we have shown that bidirectional CAGE tag clusters can identify cell type–specific enhancers in differentiated cells[39], but these were notably depleted of repetitive elements. Because we have now shown that the stem cell transcriptome is characterized by LTR usage, we sought to identify stem cell–specific enhancers residing in LTRs. We selected CAGE cluster pairs on opposite strands, separated by less than 400 bp and associated with LTR repeats, similar to Andersson *et al.*[39]. Loci in the vicinity of annotated TSSs and/or overlapping exons were removed, identifying 1,498 and 217 loci in the mouse and human data sets, respectively (**Supplementary Data Sets 1** and **2**). We noted that these loci tended to have balanced initiation (with similar expression levels on both strands) in contrast to the typically unidirectional initiation observed at annotated TSSs (**Fig. 4b** and **Supplementary Fig. 10c**), much like other enhancer regions[39]. In mouse stem cells, the top three most over-represented ERVK elements were RLTR17 (97 loci), BGLII (85 loci) and RLTR9E (53 loci), and, in human cells, LTR7 (49 loci), HERVH-int (37 loci) and LTR9 (15 loci) elements were the most abundant. The specificity

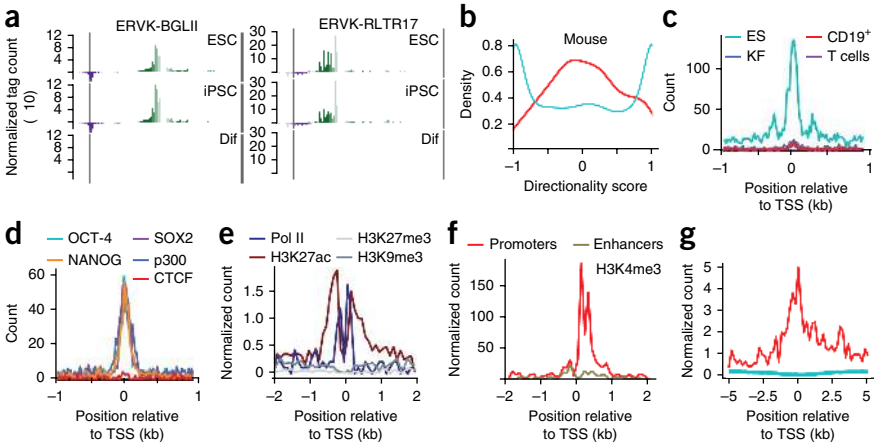

Figure 4 LTR-associated stem cell–specific regulatory elements. (**a**) Relative CAGE tag distribution from the 5′ to 3′ ends (gray vertical bars, ±10%) of mouse intergenic and intronic BGLII and RLTR17 elements. Green and purple bars indicate CAGE tags mapping to the plus and minus strands, respectively. (**b**) Density plot for the directionality score at loci showing divergent transcription overlapping with intergenic LTRs (red) or annotated TSSs (blue). Perfectly balanced transcription is reflected by a directionality score of 0: $(Exp_f - Exp_r)/(Exp_f + Exp_r)$, where Exp_f and Exp_r correspond to expression from the plus and minus strands, respectively. (**c**) Tag count density plot for mouse DNase I hypersensitivity data[5]. KF, kidney fibroblasts. (**d**) Tag count density plot for stem cell–specific transcription factor ChIP-seq data[18], the enhancer cell–associated protein p300 (ref. 18) and CTCT[5] at loci presenting divergent transcription and overlapping LTRs. (**e**) Normalized tag count density plot for mouse ChIP-seq data[5] at loci presenting divergent transcription and overlapping LTRs. Pol II, RNA polymerase II. (**f**) Normalized tag count density plot for mouse ChIP-seq data[5] for NASTs associated with LTRs and classified as promoters in **Figure 2b**, and enhancers defined as loci presenting divergent transcription and overlapping LTRs. (**g**) ChIA-PET normalized counts at loci presenting divergent transcription and overlapping LTRs (red) and at non-expressed ERVK and MaLR elements (blue).

of these potentially regulatory loci associated with mouse ERVK and human ERV1 elements is also supported by their open chromatin configuration, observed distinctly in ESCs but not in differentiated cells (**Fig. 4c** and **Supplementary Fig. 10d**) when using publicly available DNase I hypersensitivity data[5] and DNase I footprints[40]. In addition, enriched ChIP-seq[5,18] signals were found for the main stem cell–specific transcription factors (NANOG, OCT4 and SOX2) and the enhancer-related protein p300 in these putative LTR-associated regulatory regions, whereas no enrichment was found for CTCF binding (ChIP-seq data[5,41]; **Fig. 4d** and **Supplementary Fig. 10e**). The latter observation is in agreement with previous reports that some transposable elements have binding sites for stem cell–specific core transcription factors[11,41]. Finally, acetylation at lysine 27 of histone H3 (H3K27ac), reported to be associated with active enhancers[42,43], and RNA polymerase II binding signals were clearly enriched at mouse LTR-associated enhancers, unlike the repressive marks H3K9me3 and trimethylation at lysine 27 of histone H3 (H3K27me3), suggesting that we have identified active regulatory regions (ENCODE ChIP-seq data[5] for ES-Bruce4 cells; **Fig. 4e**). LTR-associated enhancers, identified on the basis of their balanced divergent transcription patterns, showed low levels of promoter-associated marks (trimethylation at lysine 4 of histone H3, H3K4me3) compared to LTR-associated NAST promoters (**Fig. 4f** and **Supplementary Fig. 10f**), supporting their classification as enhancers rather than new promoters. Lastly, using the FANTOM5 expression atlas[29], we confirmed that these putative enhancer RNAs were highly specific to stem cells (**Supplementary Fig. 10g**).

We next sought to identify target genes for these LTR-associated enhancers, analyzing RNA polymerase II–mediated interactions in mouse ESCs (in the ES-E14 cell line) by ChIA-PET (chromatin interaction

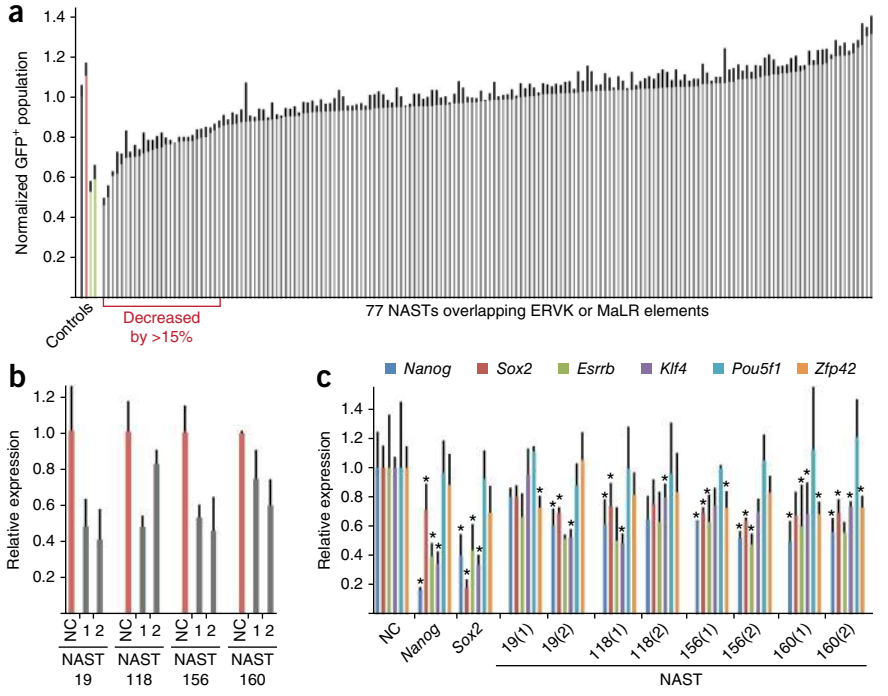

Figure 5 Implication of repeat-associated NASTs in pluripotency maintenance. (**a**) Normalized population of iPS_MEF-Ng-20D17 cells positive for GFP expression from the *Nanog* promoter, adjusted to the mock control (black), quantified by flow cytometry analysis 48 h after transient transfection with siRNA at a 20 nM concentration. The fraction of samples exceeding the threshold of a 15% decrease in *Nanog*-driven GFP expression is indicated in red. Scrambled siRNA (red) was used as a negative control, and siRNAs targeting *Nanog* and *Sox2* (green) were used as positive controls. (**b**) Knockdown efficiency measured by qRT-PCR 48 h after transfection; relative expression values were adjusted by those for cells transfected with scrambled negative control siRNA (NC; red). $n = 3$ independent experiments. Error bars, s.d. (**c**) qRT-PCR for marker genes for stemness 48 h after knockdown of NASTs. *Gapdh* levels were used for normalization, and all relative expression values were adjusted for expression levels in cells transfected with scrambled negative control siRNA. Results from *Nanog* and *Sox2* knockdown are shown for comparison. $n = 3$ independent experiments. Error bars, s.d. $^*P < 0.05$, two-sided *t* test, comparison to negative control siRNA.

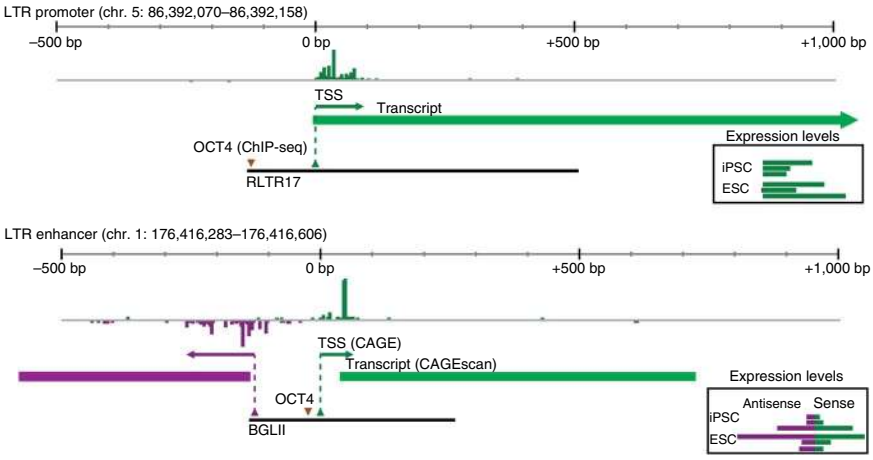

Figure 6 Genomic characteristics of the newly identified stem cell–specific transcripts. Two loci of newly identified transcripts with characteristic features of promoters (top) and enhancers (bottom) are shown as examples. CAGE-based expression levels are shown for iPSCs and ESCs (insets).

analysis by paired-end tag)[44]. Interacting loci were found to be highly enriched in mouse LTR-associated enhancers compared to non-expressed LTRs, totaling 545 interactions (**Fig. 4g**). Of these 545 interactions, 332 were intrachromosomal and occurred mainly within 100 kb of the LTR-associated enhancer (**Supplementary Fig. 10h**). Gene ontology (GO) enrichment analyses (Online Methods) for the 285 protein-coding genes physically interacting with LTR-associated enhancers identified, among other significantly enriched GO terms, chromatin organization (22 genes; adjusted $P = 5.64 \times 10^{-6}$) and the cell cycle (34 genes; adjusted $P = 3.59 \times 10^{-5}$) (**Supplementary Table 2**).

Taken together, these results show that a vast proportion of nuclear LTR-derived transcripts originate from distal regulatory regions likely implicated in chromatin organization and cell cycle regulation.

Implication of LTR transcripts in pluripotency maintenance

To test the putative implication of LTR-associated NASTs in the genetic regulation of pluripotency, we sought to perturb the expression of candidate elements using transient knockdown with small interfering RNA (siRNA) in mouse iPSCs carrying a *GFP* reporter gene under the control of the *Nanog* promoter (iPS_MEF-Ng-20D17 cells)[21]. NANOG has been reported to uniquely mark the pluripotent state[45,46]. We targeted 77 LTR-associated NAST candidates (**Supplementary Table 3**) chosen from rather highly expressed examples with a median nuclear expression value of 15.6 tpm (minimum of 0.8 tpm, maximum of 508.9 tpm, from mean expression values for the 6 mouse stem cell nuclear samples). The large majority of these candidates (64/77) carried promoter-associated histone marks. In an initial screening experiment, a decrease in the GFP-positive population of greater than 15% (with this threshold based on the results for multiple negative controls; **Supplementary Fig. 11** and **Supplementary Table 4**), measured by flow cytometry analysis, was observed for 25 tested NASTs (**Fig. 5a**). Among these candidates, four NASTs could reproducibly be knocked down by two different siRNAs (**Fig. 5b**). The loss of stemness caused by perturbation of these four NAST candidates was further confirmed by decreased expression of multiple marker genes for stemness (**Fig. 5c**).

In conclusion, our knockdown experiments identified several NAST candidates whose perturbation resulted in the downregulation of multiple marker genes for stemness. These findings suggest a direct role for some mouse NASTs, originating from ERVK and MaLR elements, in the genetic regulatory network for the maintenance of pluripotency. Notably, NASTs are likely to be implicated in numerous physiological pathways and to have many phenotypic classes not necessarily related to pluripotency maintenance, such as cell proliferation or cell adhesion. NASTs should thus not be expected to behave as a single functional class. Further work will be needed to decipher their molecular mechanisms of action and to identify the full set of NASTs with other phenotypes related to cell proliferation, cell adhesion or the naive state.

DISCUSSION

To our knowledge, this study provides the most comprehensive transcriptional profiling of mouse and human stem cells, based on four complementary high-throughput sequencing methods. Characterizing the deepest part of the transcriptome for each specific subcellular compartment is essential to an understanding of the complexity of previously unannotated RNAs, both in human and mouse stem cells, with potential regulatory features, such as the NASTs described in this study. The identification of promoter and enhancer histone marks on a substantial fraction of the NASTs, which map either in an antisense orientation relative to annotated genes or reside in intronic and intergenic regions, and the specificity of CAGE in identifying new 5′ ends together suggest that these TSSs are products of new transcription rather than processed RNA. Analysis of the syntenic conservation of differentially expressed clusters showed that a large fraction of the NASTs are species specific and overlap with repetitive genomic regions. The findings summarized above support the notion that wide use of stem cell–specific LTR-derived promoters controls the expression of new nuclear transcripts—potentially new lncRNAs—on a larger scale than previously thought. We report here 639 human LTR-associated NASTs, of which only 39 and 12 overlap with previously described LTR-associated lncRNAs[13] and LTR-associated very long lncRNAs[47], respectively. In addition, 2,372 NASTs originating in LTRs were identified in the mouse, thereby enriching knowledge of active LTR-derived promoters in this species, as only 8.8% of these were previously reported in the Guttman *et al.* lncRNA catalog[28].

The relatively low conservation of NASTs could potentially reflect the different biological properties of human and mouse ESCs and iPSCs. Indeed, human ESCs have been described as corresponding more with mouse-derived epiblast stem cells than with mouse ESCs[48,49]. However, it seems striking that different LTR subfamilies have been recruited independently within the stem cell regulatory networks in humans and mice. Similar to what has recently been shown for CTCF binding sites[41], one could speculate that the retrotransposition of LTRs and, more specifically, human ERV1 and mouse ERVK elements has participated in the generation of stem cell–specific promoters throughout the corresponding genomes. Past studies have shown that species-specific transposable elements, human ERV1 and mouse ERVK, have expanded chromatin-binding sites for two main stem cell transcription factors, NANOG and OCT4, in ESCs[12], but the relevance of these sites was not fully understood. In addition, mouse ERVK elements are also known to provide binding sites for SOX2 (ref. 11). Our analysis suggests that LTR-derived NASTs are directly under the control of the main stem cell–specific transcription factors.

Hypothetically, once a repetitive element had acquired transcription factor binding sites by mutational drift, colonization by its genome and subsequent positive evolutionary selection would fit with our observation of widespread stem cell–specific promoters associated with a few specific repetitive elements. In line with this hypothesis, a recent study[50] reported the association of tissue-specific enhancers with transposable elements. One could thus speculate that many more biologically functional LTR-derived transcripts might be implicated in stem cell–specific processes, similar to the iPSC-enriched lincRNA-RoR, whose expression is driven from an LTR7-derived promoter and that has a role in stem cell survival, possibly by promoting cellular stress pathways[15]. Our study provides another level of comprehension by showing that RNA is actively transcribed from these LTR-derived loci (**Fig. 6**) and that some of these LTR-associated transcripts participate in the regulatory network of pluripotency maintenance.

On the basis of their divergent, balanced transcription patterns, we also find that many LTR-derived transcripts originate from putative enhancers (**Fig. 6**). These putative enhancers physically interact with the promoters of genes implicated in chromatin state and cell cycle regulation. Remarkably, in their systematic characterization of active enhancers across all tissues and primary cell samples from the FANTOM5 expression atlas[29], Andersson *et al.*[39] found that somatic cell–specific enhancers are generally depleted of repeat elements. In contrast, genome-wide DNA methylation data identified cell type–specific LTR-associated enhancers in somatic tissues[50].

This study, together with recent reports, has probably just begun to unravel the set of unexpected functions of retrotransposons in stem cell biology.

URLs. Vegan Community Ecology Package, http://CRAN.R-project.org/package=vegan.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All sequencing data have been deposited at the DNA Data Bank of Japan (DDBJ) under accession DRA000914.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank the RIKEN GeNAS sequencing platform for sequencing of the libraries. This work was supported by a grant to P.C. from the Japan Society for the Promotion of Science (JSPS) through the Funding Program for Next-Generation World-Leading Researchers (NEXT) initiated by the Council for Science and Technology Policy (CSTP), by a grand-in-aid for scientific research from JSPS to P.C. and A.F., and by a research grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to the RIKEN Center for Life Science Technologies. FANTOM5 was made possible by a research grant for the RIKEN Omics Science Center from MEXT Japan to Y. Hayashizaki and by a grant for Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from MEXT to Y. Hayashizaki. A.F. was supported by a JSPS long-term fellowship (P10782) and by a Swiss National Science Foundation Fellowship for Advanced Researchers (PA00P3_142122). K.H. was supported by European Union Framework Programme 7 (MODHEP project) for P.C. A.B. was supported by the Sigrid Juselius Foundation Fellowship. D.Y. and H.K. were supported by the Japan Science and Technology Agency CREST. R.A. and A. Sandelin were supported by funds from FP7/2007-2013/ERC grant agreement 204135, the Novo Nordisk Foundation, the Lundbeck Foundation and the Danish Cancer Society.

AUTHOR CONTRIBUTIONS

P.C. led the project and oversaw the analyses. P.C., Y.H. and A.R.R.F. contributed to the design of the study. A.F., D.Y., M.S., C.A.K., A. Saxena, A.B., H.S., H.K., Y.N. and Y.H. contributed to data generation. A.F., K.H., I.V., N.B., M.d.H., A.R.R.F., A.K., R.A. and A. Sandelin contributed to data processing and analyses. C.-H.W. and C.-L.W. produced and analyzed ChIA-PET data. A.F., A.B., A.R.R.F. and P.C. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
2. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
3. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
4. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
7. Faulkner, G.J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).
8. Santoni, F.A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**, 111 (2012).
9. Peaston, A.E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
10. Macfarlan, T.S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
11. Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
12. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
13. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell–specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
14. Kolle, G. *et al.* Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. *Genome Res.* **21**, 2014–2025 (2011).
15. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113–1117 (2010).
16. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).
17. Ng, S.Y., Johnson, R. & Stanton, L.W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* **31**, 522–533 (2012).
18. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).
19. Wang, Y. *et al.* Embryonic stem cell–specific microRNAs regulate the G1-S transition and promote rapid proliferation. *Nat. Genet.* **40**, 1478–1483 (2008).
20. Melton, C., Judson, R.L. & Blelloch, R. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature* **463**, 621–626 (2010).
21. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (2007).
22. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).
23. Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* **7**, 528–534 (2010).
24. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
25. Efroni, S. *et al.* Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell* **2**, 437–447 (2008).
26. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
27. 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).
28. Guttman, M. *et al. Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
29. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
30. Sigova, A.A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **110**, 2876–2881 (2013).
31. Min, I.M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25**, 742–754 (2011).
32. Hinrichs, A.S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

33. Rebollo, R., Romanish, M.T. & Mager, D.L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* **46**, 21–42 (2012).
34. Rowe, H.M. & Trono, D. Dynamic control of endogenous retroviruses during development. *Virology* **411**, 273–287 (2011).
35. Karimi, M.M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8**, 676–687 (2011).
36. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
37. Rowe, H.M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
38. Kim, T.K. *et al.* Widespread transcription at neuronal activity–regulated enhancers. *Nature* **465**, 182–187 (2010).
39. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
40. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
41. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
42. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
43. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
44. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
45. Chambers, I. *et al.* Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643–655 (2003).
46. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642 (2003).
47. St Laurent, G. *et al.* VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol.* **14**, R73 (2013).
48. Brons, I.G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
49. Tesar, P.J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
50. Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).

ONLINE METHODS

Cell culture. Mouse iPS_MEF-Ng-20D17 cells (miPS.F)[21], mouse ESCs (B6G-2 (ref. 51), FVB-1 (ref. 52) and Nanog^(βgeo/+)ES (mESR08, CIRA)) and mouse embryonic fibroblasts (MEFs; MEF_Ng-20D17) were purchased from the RIKEN BioResource Center. All RIKEN BioResource Center cell lines are regularly authenticated and tested for the absence of mycoplasma. Human fetal dermal fibroblasts (HDF-f) were purchased from Cell Applications.

miPS.F, mESB6G-2, mESFVB-1 and mESR08 cells were grown under feeder-free conditions in mouse ESC medium containing DMEM (Wako), 1,000 U/ml leukemia inhibitory factor (LIF; Millipore), 15% FBS (Gibco), 2.4 mM L-glutamine (Invitrogen), 0.1 mM non-essential amino acids (NEAA; Invitrogen), 0.1 mM 2-mercaptoethanol (Gibco), 50 U/ml penicillin and 50 μg/ml streptomycin (Gibco). Culture media were changed daily, and cells were passaged every 2–3 d.

Established mouse iPSCs were cultured on MEFs treated with mitomycin (Sigma) in DMEM containing 20% FBS, 2,000 U/ml LIF, 1% NEAA, 0.1 mM 2-mercaptoethanol, 2.4 mM L-glutamine and three inhibitors (3i)[53].

MEFs and HDF-f cells were cultured in DMEM containing 20% FBS, 50 U/ml penicillin and 50 μg/ml streptomycin.

Human ESCs (KhES-1, KhES-2 and KhES-3 cells) were cultured on mitomycin-treated MEF feeder cells in Primate ES medium (ReproCELL) supplemented with 5 ng/ml basal fibroblast growth factor (bFGF; Wako), in accordance with institutional and national regulations (under RIKEN Yokohama ethics application H20-2(8)).

Lymphocyte isolation. Human and mouse $CD19^+$ cells (B lymphocytes) and $CD3^+$ cells (T lymphocytes) were isolated from fresh blood and spleen, respectively, using MACS beads (Miltenyi Biotech) and were cultured in lymphocyte complete medium (RPMI1640, Sigma) containing 10% FBS, 5 ng/ml interleukin (IL)-4 and 25 mg/ml lipopolysaccharide (LPS) for B cells or Dynabeads T-Activator CD3/CD28 (Veritas) for T cells.

iPSC derivation. Two iPSC lines were generated in this study from mouse primary T lymphocytes (miPS.T cells; clone i103 H12) and from human fibroblasts (HDF-f, hiPS.F cells; clone hi6) (**Supplementary Fig. 1**).

Retrovirus preparation was carried out as previously described[54]. Mouse T cells were infected at 1×10^5 cells/ml in the presence of 10 mg/ml polybrene (Sigma), 5 ng/ml IL-4 (R&D Systems) and 25 mg/ml LPS (Sigma). After 24 h, medium was replaced with lymphocyte complete medium, and cells were seeded on mitomycin-treated MEF feeder cells. Seventy-two hours after transduction, medium was replaced with mouse ESC medium, and medium was changed every other day until ESC-like colonies formed. Colonies were isolated, dissociated with trypsin (Invitrogen) and transferred to stem cell medium (DS Pharma Biomedical) maintained with 2,000 U/ml LIF and 0.1 mM 2-mercaptoethanol and kept in culture for further experiments.

hiPS.F cells were derived using a non-integrating Sendai-based viral vector (SeV) coding for the four Yamanaka factors, following the methods detailed in Fusaki *et al.*[55].

Extraction of nucleus-enriched and cytoplasmic RNAs. For all cell lines (**Table 1**), nucleus-enriched and cytoplasmic RNA fractions were isolated from 5 to 10 million cells. Cells were first lysed in chilled lysis buffer (0.8 M sucrose, 150 mM KCl, 5 mM $MgCl_2$, 6 mM 2-mercaptoethanol and 0.5% NP-40) and spun for 5 min at 10,000*g* (4 °C). Supernatants containing cytoplasmic fractions were collected and promptly mixed with three volumes of TRIzol-LS Reagent (Life Technologies). Nuclei pellets were washed twice with lysis buffer before resuspension in TRIzol Reagent. A miRNEasy kit (Qiagen) was used according to the manufacturer's protocol to extract both nucleus-enriched and cytoplasmic RNA fractions. During the RNA purification process, samples were treated with DNase I (Qiagen).

Library preparation. CAGE libraries were prepared starting with 0.5 to 5 μg of RNA, following the protocols developed in our laboratory[22,56].

CAGEscan libraries were prepared as described[23], starting with 50 ng of RNA treated with T4 polynucleotide kinase (New England Biolabs) before digestion with terminator 5′ phosphate–dependent exonuclease (Epicentre Biotechnologies).

Short RNA-seq libraries were prepared (TruSeq_Small_RNA_Sample_V2.0, Illumina) from 80 to 500 ng of RNA. To not limit the sequencing to the miRNA size range, short RNA-seq libraries were size selected in fractions containing inserts from 15 to 40 bp and 80 to 280 bp in length.

Before RNA-seq library preparation (ScriptSeq_v2_RNA-Seq, Epicentre Biotechnologies), mESR08 and mESB6G2 cytoplasmic RNA samples were treated using the Ribo-Zero Magnetic kit (Epicentre Biotechnologies); other samples were treated with the Ribo-Zero rRNA Removal Kit Low Input (Epicentre Biotechnologies).

CAGE processing, analyses and CAGE cluster annotation. CAGE libraries were sequenced on the Illumina HiSeq 2000 platform with a read length of 50 bases. After discarding sequences with ambiguous base calling, splitting sample reads by barcodes and removing linker sequences and artifactual linker adaptor sequences (using TagDust[57]), reads were of 26 to 42 bases in length. CAGE reads were mapped to hg19/GRCh37 and mm9/NCBI37 using Burrows-Wheeler Aligner (BWA) v0.5.6 (ref. 58). Only reads with MapQ values over 10 and therefore mapping to single loci in the genomes were used in our analyses. Subsequently, reads mapping to ribosomal DNA were eliminated.

CAGE tag 5′ genomic coordinates were used as input for Paraclu[59] clustering with the following parameters: (i) a minimum of 5 tags per cluster, (ii) maximum density/baseline density ≥ 2 and (iii) a maximal cluster length of 200 bp.

Estimation of the richness score on the basis of the expression of CAGE clusters, shown in **Figure 1b**, was calculated using the Vegan R package.

Annotations of CAGE tag clusters (**Fig. 1c**) were based on the GENCODEV10 (ref. 60), RefSeq[61], UCSC KnownGenes[62], lincRNA transcript[63] and H-inv 7.0 (ref. 64) databases for humans, and the RefSeq, Ensembl[65] and UCSC KnownGenes databases (retrieved from the UCSC browser in January 2012) were used for mice. Hierarchical multiple annotation of CAGE tag clusters was performed starting with (i) sense TSSs and exons, (ii) antisense TSSs and exons, (iii) introns, (iv) sequence ±1 kb relative to TSSs and (v) intergenic sequences. Repetitive element annotations were retrieved from the UCSC browser, which ran RepeatMasker[36] version open-3-2-7 using the program by A. Smit, with sensitive settings on the 20090120 release of the Repbase Update library of repeats from the Genetic Information Research Institute.

Histone mark–based classification of the CAGE clusters associated with NASTs (**Fig. 2b** and **Supplementary Fig. 6**) was performed using ChIP-seq data[5] for human H1-ESC and mouse ES-Bruce-4 and ES-E14 cells. Loci carrying a stronger signal for monomethylation at lysine 4 of histone H3 (H3K4me1) than for H3K4me3 and carrying H3K27ac were classified as enhancers[42,66], whereas clusters with stronger signal for H3K4me3 than for H3K4me1 and/or carrying H3K9ac marks were considered to be promoters. CAGE clusters carrying H3K9me3 and/or H3K27me3 marks were annotated as repressed. Finally, CAGE clusters presenting trimethylation at lysine 36 of histone H3 (H3K36me3) were annotated as gene body.

CAGEscan processing. CAGEscan libraries were sequenced on the Illumina HiSeq 2000 platform with a paired-end read length of 50 bases. After discarding sequences with ambiguous base calling (identified as N), splitting sample reads by barcodes and removing linker sequences and artifactual linker adaptor sequences (using TagDust[57]), read lengths were 36 bases for 5′ reads and 50 bases for 3′ reads. Subsequently, reads mapping to ribosomal DNA were eliminated. 5′ and 3′ reads were mapped independently to hg19/GRCh37 and mm9/NCBI37 using BWA v0.5.6 (ref. 58).

CAGEscan assemblies were performed independently for each library using CAGE clusters as a guide, following previously described methods[23] and using properly paired reads with combined MapQ values of greater than 50.

Short RNA-seq processing and analyses. Short RNA-seq libraries were sequenced on the Illumina HiSeq 2000 platform with read lengths of 50 (short fraction, 15 to 40 nt) and 100 (long fraction, 80 to 280 nt). After splitting sample reads by barcode, removing linker sequences, discarding sequences with ambiguous base calling (identified as N) and eliminating reads mapping to ribosomal DNA, lengths were 16–51 bases for the shorter fraction and 16–101 bases for the longer fraction. Tags for the long fraction were mapped by BWA v0.5.6 (ref. 58), and reads were mapped with Delve[6] for the short fraction.

Resulting tags from nuclear and cytoplasmic samples were pooled and clustered with Paraclu[59], using the following parameters: (i) a minimum of 30 tags per cluster, (ii) maximum density/baseline density ≥ 2 and (iii) maximal cluster length of 100 bp.

RNA-seq processing and analyses. RNA-seq libraries were sequenced on the Illumina HiSeq 2000 platform with 100-nt paired-end reads. After splitting sample reads by barcode, eliminating reads mapping to ribosomal DNA and discarding sequences with ambiguous base calling (identified as N), properly paired sequences were mapped to genomes using TopHat v1.4.1 (ref. 67), and transcript assemblies were carried out for each sample separately with Cufflinks v1.3.0 (ref. 26) using Ensembl[65] transcripts as a guide. Resulting transcripts from RNA-seq data in nuclear and cytoplasmic samples were merged separately using Cuffmerge[68].

Transcript copy number per cell. Total RNA was extracted (RNeasy kit, Qiagen) from 5 million cells (iPS_MEF-Ng-20D17 and mESR08) in biological triplicate, with average yields of 49.9 µg (±6.7 µg). Firefly reference RNA was prepared by *in vitro* transcription (mMESSAGE mMACHINE T7 kit, Ambion) from a pcDNA3.1 plasmid (Invitrogen) including the firefly luciferase cDNA sequence. Reverse transcription was performed using random hexamers (PrimeScript First-Strand cDNA Synthesis kit, TAKARA) with 1 µg of total RNA (equivalent of 1×10^5 cells), and samples were spiked with 1×10^5 or 1×10^6 firefly reference RNA molecules. C_t values were obtained for the reference firefly RNA, NASTs and *Gapdh* (primers listed in **Supplementary Table 5**) using SDS version 2.1 software (Applied Biosystems).

Chromatin interaction analysis by paired-end tag. The ChIA-PET data used in the present study are described in Zhang *et al.*[44].

GO term enrichment analysis (**Supplementary Table 2**) was performed using the WebGestalt tool[69] with the 285 genes interacting with LTR-associated enhancers used as input and Entrez Gene[70] protein-coding genes set as background. The *P* values obtained were adjusted by Bonferroni correction, and the significance threshold was set at 0.01.

siRNA transfection assays. iPS_MEF-Ng-20D17 cells[21] were cultured, from 24 h before transfection until the end of the experiment, in ESC medium containing 50 U/ml LIF, therefore maintaining their pluripotent state with minimum activation of the LIF pathway[71]. We seeded 30,000 cells/well in 12-well plates or 20,000 cells/well in 24-well plates 24 h before transfection. In ESC medium depleted of antibiotic, 20 nM of siRNA (Stealth RNAi siRNA, Life Technologies; **Supplementary Table 3**) was transfected into cells using Lipofectamine RNAiMAX reagent (Life Technologies), following the manufacturer's instructions. siRNAs targeting non-expressed repeat elements as well as non-expressed genes with promoters overlapping LTR elements were used as negative controls (**Supplementary Table 4**) in addition to commercially available scrambled siRNA (Negative Control, Medium GC duplexes 1 and 2, Life Technologies) and siRNA targeting the luciferase gene (Life Technologies). siRNAs targeting *Nanog* and *Sox2* were used as positive controls (**Supplementary Table 4**).

Forty-eight hours after transfection, iPS_MEF-Ng-20D17 cells were collected from biological replicates and processed for flow cytometry analysis, using a BD FACSAria II instrument. Cells positive for *Nanog*-driven GFP expression (gate FITC-A > 10^4) were quantified and normalized to the mock condition.

Knockdown efficiency was measured by qRT-PCR (primers are listed in **Supplementary Table 5**). Total RNA (1 µg), extracted with the RNeasy kit following the manufacturer's protocol, was reverse transcribed using random hexamers (PrimeScript First-Strand cDNA Synthesis kit). cDNA synthesized from cytoplasmic RNA was used to assess the expression of stemness marker genes (*Nanog*, *Sox2*, *Esrrb*, *Klf4*, *Pou5f1* (also known as *Oct4*) and *Zfp42* (also known as *Rex1*)). C_t values were obtained using SDS version 2.1 software. Relative RNA levels were calculated using the $\Delta\Delta C_t$ method[72], with expression normalized to that of *Gapdh*.

51. Shimizukawa, R. *et al.* Establishment of a new embryonic stem cell line derived from C57BL/6 mouse expressing EGFP ubiquitously. *Genesis* **42**, 47–52 (2005).
52. Wakayama, T. *et al.* Differentiation of embryonic stem cell lines generated from adult somatic cells by nuclear transfer. *Science* **292**, 740–743 (2001).
53. Ying, Q.L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
54. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
55. Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci.* **85**, 348–362 (2009).
56. Salimullah, M., Sakai, M., Plessy, C. & Carninci, P. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* **2011**, pdb.prot5559 (2011).
57. Lassmann, T., Hayashizaki, Y. & Daub, C.O. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839–2840 (2009).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Frith, M.C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
60. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
61. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
62. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).
63. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
64. Yamasaki, C. *et al.* H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res.* **38**, D626–D632 (2010).
65. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
66. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
67. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
68. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
69. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33**, W741–W748 (2005).
70. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58 (2005).
71. Hasegawa, Y. *et al.* CC chemokine ligand 2 and leukemia inhibitory factor cooperatively promote pluripotency in mouse induced pluripotent cells. *Stem Cells* **29**, 1196–1205 (2011).
72. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).

doi:10.1038/ng.2965