Citation for published version:
Wang, J, Xie, G, Singh, M, Tashakkori Ghanbarian, A, Raskó, T, Szvetnik, A, Cai, H, Besser, D, Prigione, A, Fuchs, NV, Schumann, GG, Chen, W, Lorincz, MC, Ivics, Z, Hurst, LD & Izsvák, Z 2014, 'Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells', Nature, vol. 516, no. 7531, pp. 405-409. https://doi.org/10.1038/nature13804

DOI:
10.1038/nature13804

Publication date:
2014

Document Version
Peer reviewed version

Link to publication

Primate-specific endogenous retrovirus driven transcription defines naïve-like stem cells

Jichang Wang[1&], Gangcai Xie[1,2&], Manvendra Singh[1], Avazeh T. Ghanbarian[3], Tamás Raskó[1], Attila Szvetnik[1], Huiqiang Cai[1], Daniel Besser[1], Alessandro Prigione[1], Nina V. Fuchs[1,4], Gerald G. Schumann[4], Wei Chen[1], Matthew C. Lorincz[5], Zoltán Ivics[4], Laurence D. Hurst[3*], Zsuzsanna Izsvák[1*]

[1] Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Strasse 10, 13125 Berlin, Germany.

[2] Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, 320 Yue Yang Road, Shanghai 200031, China.

[3] University of Bath, Department of Biology and Biochemistry, Bath, Somerset, UK, BA2 7AY.

[4] Paul-Ehrlich-Institute, Division of Medical Biotechnology, Paul-Ehrlich-Strasse 51-59, 63225 Langen, Germany.

[5]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada V6T 1Z3.

[&] equal contribution

*Corresponding authors

For correspondence:

Zsuzsanna Izsvák
Max Delbrück Center for Molecular Medicine
Robert Rössle Strasse 10, 13092 Berlin, Germany
Telefon: +49 030-9406-3510
Fax: +49 030-9406-2547
email: zizsvak@mdc-berlin.de
http://www.mdcberlin.de/en/research/research_teams/mobile_dna/index.html

and

Laurence D. Hurst
Professor of Evolutionary Genetics
Department of Biology and Biochemistry, University of Bath
Bath, Somerset, UK BA2 7AY
tel: +44 (0)1225 386424
Fax: fax: +44 (0)1225 386779
email: l.d.hurst@bath.ac.uk
http://people.bath.ac.uk/bssldh/LaurenceDHurst/Home.html

Naïve embryonic stem cells (ESCs) hold great promise for research and therapeutics as they have broad and robust developmental potential. While such cells are readily derived from mouse blastocysts it has been impossible to easily isolate human equivalents[1,2], although human naïve-like cells have been artificially generated (rather than extracted) by coercion of human primed ES cells by modifying culture conditions[2-4] or through transgenic modification[5]. Here we show that a sub-population within cultures of human ESCs (hESCs) and induced pluripotent stem cells (hiPSCs) manifest key properties of naïve state cells. These naïve-like cells can be genetically tagged, and are associated with elevated transcription of HERVH, a primate-specific endogenous retrovirus (ERV). HERVH elements provide functional binding sites for a combination of naïve pluripotency transcription factors, including LBP9, recently recognized as relevant to naivety in mice[6]. LBP9/HERVH drives hESC-specific alternative and chimeric transcripts, including pluripotency modulating long non-coding RNAs (lncRNAs). Disruption of LBP9, HERVH and HERVH-derived transcripts compromises self-renewal. These observations define HERVH expression as a hallmark of naïve-like hESCs, and establish novel primate-specific transcriptional circuitry regulating pluripotency.

While many genes are involved in pluripotency, transposable element (TE) transcription, particularly involving ERVs, has wired different genes into the network in humans and mice[7]. Given a role for ERVs in pluripotency[8-10], we surveyed RNAseq data of human pluripotent stem cells (hPSCs), notably hESCs and hiPSCs finding that several TEs are expressed at higher levels in hPSCs, ERV1 type of long terminal repeat (LTR) retroelements being foremost, of which HERVH was the most highly expressed[8,11] (Figs. 1a-b, E1a-b). Uniquely aligned reads (Table S6) indicate that 550 of the 1225 full-length HERVH genomic copies are transcribed in hPSCs (Figs. E1c-d; Table S7). Raised transcription was associated with elements containing consensus LTR7 rather than diverged variants (LTR7B/C/Y: Table S7). Lower expression of other ERVs (Fig. 1b) was confirmed via qRT-PCR (Fig. 1c). We focused on HERVH as this was the only one detected by qRT-PCR in all hiPSC lines analysed (Fig. 1c). Results are robust to use of reads that map to more than one location (Table S16).

To address how specific HERVH transcription is to hPSCs we compared RNAseq datasets of hPSCs and multiple differentiated cells and tissues (Fig. E1c; Tables S4, S5, S7). In agreement with our hiPSC data, HERVH transcription was highest in hPSC lines. The majority of the transcribed loci are identical between hiPSCs and hESCs (Figs. E1c-d). HERVH transcription levels are much lower in both differentiated cells and cancer cell lines (Fig. E1c).

HERVH transcription levels are higher in hiPSCs at early passages following reprogramming (Fig. 1d), indicating that the reprogramming process itself might induce HERVH expression. At later passages the transcription of HERVH in hiPSCs approaches hESC levels.

Consistent with HERVH transcription in hPSCs, ChIP-seq data show that, in contrast to HERVK and inactive HERVHs, active HERVHs are marked with transcriptionally active histone marks[11,12] (H3K4me1/2/3, H3K9ac, H3K36me3 and H3K79me2), while the repressive marks (H3K9me3 and H3K27me3) are rare, indicating functioning as active promoter/enhancers (Figs. 2a, E2a-e). Notably, active HERVHs are also enriched with binding sites of the pluripotency regulators/modifiers CHD1[13] and Myc/Max[14] (Figs. E2b-c; Table S15). HERVH activation is also inversely correlated with the DNA methylation status of LTR7 of HERVH, as evidenced by hypomethylation in active LTR7 regions in hPSCs[15] (Fig. E2f).

To determine whether HERVH is a direct target of core pluripotency–associated transcription factors (TFs) we interrogated HERVH in hESC_H1 ChIP-Seq data[3]. This identified NANOG and OCT4 (Fig. E3a). A candidate KLF4 binding site was also identified within HERVH's LTR (Fig. 2b). We additionally asked which TF motifs are significantly enriched across four *in silico* tests (Fig E3b). Only one, LTR-binding protein 9 (LBP9) – alias murine Tfcp2l1 - was significant across all analyses (Fig. E3b). Tfcp2l1 is within the Oct4 interactome[16] and binds regulatory regions of Oct4 and Nanog[17] in mESCs. LBP9's direct binding to LTR7 is confirmed by ChIP-qPCR and EMSA (Fig. 2c, and Fig. E3c). EMSA further demonstrates LBP9/NANOG cooperation in binding LTR7 (Fig. E3c), consistent with synergy following simultaneous over-expression (Fig. E7c). LBP9-specific binding was also detected in the 5'-region of NANOG (Fig. 2c).

In vitro differentiation assays show that HERVH transcription levels decline over time in parallel with declines in OCT4, NANOG and LBP9 (Fig. E3d), suggesting a role in HERVH expression. As expected, ectopic expression of LBP9, OCT4, NANOG and KLF4 activated the pT2-LTR7-GFP#2 reporter and enhanced endogenous HERVH transcription levels in human primary fibroblast (HFF-1), while overexpression of c-MYC or SOX2 had no effect (Fig. 2d, E7c). Conversely, a complementary 'loss of function' RNAi assay in hESCs_H9 revealed that HERVH transcription levels were reduced following OCT4, NANOG and LBP9, but not SOX2, knockdown (KD) (Figs. 2e-f).

We confirmed that LBP9 directly stimulates HERVH-driven expression, by comparing in hiPSCs signals of a wild-type (WT) pT2-LTR7-GFP#1 reporter construct and a mutant lacking the LBP9 motif (ΔLBP9: Fig. E7d). When WT and mutant constructs were transfected into hiPSCs, the GFP signal was clearly detected from the WT reporter, but it was decreased by 2-fold in ΔLBP9 (Fig. E7d).

ESC-specific TFs OCT4, NANOG, KLF4 and LBP9 thus drive transcription in hPSCs. In contrast to mice in which LBP9 binding sites are genomically distinct from those other pluripotency TFs[6], the key pluripotent TFs cluster within the primate-specific HERVH (Fig. 2b).

To test the functional importance of HERVH, we analysed RNAseq data to investigate the influence of LTR7/HERVH on the expression of neighbouring regions. We find that LTR7 initiates chimeric transcripts, functions as an alternative promoter or modulates RNA processing from a distance (Figs. 3a, E4b; Tables S8-9). 128 and 145 chimeric transcripts were identified in hiPSCs and hESCs, respectively (Fig. E4a; Tables S8,-9). One gene can contribute to multiple chimeric transcripts. The chimeric transcripts between HERVH and a downstream gene generally lack the 5' exon(s) of the canonical version (e.g. SCGB3A2) while part of HERVH/LTR7 is exonized (e.g. RPL39L) (Fig 3a). A significant fraction of HERVH sequence can be incorporated into novel, lineage-specific genes (e.g. ESRG: Fig. 3a) or lncRNAs (e.g. RP11-69I8.2: Fig. E4d and Table S10). We confirmed several hPSC specific chimeric transcripts by RT-PCR (Fig. 3a). Transcriptional start signals commonly map to HERVH-LTR boundary regions (Fig. E4c). Unlike the chimeric transcripts the canonical genes are commonly not expressed in pluripotent cells.

Nearly 10% of the transcripts driven off HERVH are annotated as lncRNA[12] (see Table S11 for coding potential). 54 transcripts were identified that are commonly detected in hPSCs, while the rest were sporadic (Fig. E4d). The former set includes linc-ROR and linc00458, known to modulate pluripotency[18,19]. Alignment of the 22 most highly expressed transcripts reveals an LTR7/HERVH-derived conserved core domain (CD) (Fig. E4f). The domain is predicted to bind RNA-binding proteins, including pluripotency factors (e.g. NANOG) and pluripotency-associated histone modifiers (e.g. SET1A and SETDB1) (Fig. E4g). In agreement with a role in pluripotency, linc00458 physically interacts with SOX2[19].

To explore the effect of either LBP9 or specific HERVH-derived transcripts on the reprogramming process, we asked whether forced expression of LBP9, ESRG or the conserved domain of lncRNAs (LTR7-CD) modulates the fibroblast-hiPSC transition. While the overexpressed gene products affect neither pluripotency nor self-renewal (Figs. E5a-b), all facilitate reprogramming by accelerating the mesenchymal-epithelium transition or hiPSC maturation (Figs 3b, E5c).

While LBP9 is key to the murine naïve state[6,20], HERVH is primate-specific. To determine whether HERVH/LBP9 delineates a primate-specific pluripotency circuitry, we performed "loss of function" experiments using small hairpin RNAs (shRNAs) against LBP9 or HERVH (Figs. 3c-f, E5d-g). Pluripotency-associated TFs and markers are down-regulated, while multi-lineage differentiation markers are up-regulated upon knockdown of either, but not in controls (Fig. 3c-d, E5f-g). Depletion of LBP9 or HERVH in hESCs thus results in loss of self-renewal. Knockout of LBP9 similarly abolishes hESC self-renewal (Figs. E5h-j). In contrast to hPSCs, the Tfcp2l1/LBP9 knockdown in mESCs does not reduce levels of Oct4, Sox2 and Nanog in serum-based conditions (Fig. E5k)[21], but only in 2i[6]. In fact, Tfcp2l1/LBP9 does not affect self-renewal, but rather differentiation potential (Fig. E5k).

Genome-wide gene expression patterns are highly similar between LBP9 and HERVH knockdowns (Fig. 3e), consistent with LBP9 regulating HERVH-driven expression. 1094 of the 2627 genes are similarly regulated in LBP9/HERVH knockdowns (Fig. 3f; Table S12). While some HERVH-derived chimeric transcripts are potentially directly affected by depletion of HERVH (Tables S13-14), qRT-PCR identifies 19 HERVH-derived lncRNAs, down-regulated in response to both HERVH and LBP9 knockdowns (Fig. E4e).

While several of the differentially expressed genes are associated with murine pluripotency, the LBP9/HERVH-driven list of transcripts defines a primate-specific pluripotency network. Our analyses defined two classes of genes, (I) those conserved between mouse and human that contribute to the pluripotency in both, and (II) a primate-specific group that includes (a) those with an orthologous partner, but are not involved in murine pluripotency and (b) novel (not in mouse) transcripts (Figs. E4b, E4d). Several HERVH elements in class IIa affect gene expression in *cis*, and drive specific genic isoforms (e.g. SCGB3A2). A subset of class IIb contains HERVH-derived novel sequences (e.g. linc-ROR, linc000548, ESRG) (Fig. E4d).

We examined one class IIb transcript in detail. ESRG has a putative open reading frame (ORF) only in human (Fig. E6a; Supplementary Data 1), and is uniquely expressed in human inner cell mass (ICM) and PSCs (Fig. E6b). Knockdown of ESRG compromised self-renewal of hESCs, as many pluripotency-associated genes were decreased, while SOX2 expression was slightly elevated (Figs. E6c-e). The KD-ESRG colonies lost their hESC morphologies and committed to differentiation (Figs. E6e-f). Expression of ESRG along with the OSKM pluripotency factors has a similar effect on the reprogramming process compared with LBP9 (Fig E5c). ESRG is thus an HERVH-associated novel gene required for human-specific pluripotency, with a more specific phenotype than upstream regulators.

Given that the naïve-associated TFs together cluster on HERVH and the HERVH-derived products are essential for primate pluripotency, we asked whether

HERVH-driven transcription marks the naïve-like stage in hPSC cultures. To explore this the reporter construct, pT2-LTR7-GFP#2 was integrated into the genome of either mouse or human PSCs (Figs. 4a, E7a-b, E8i) by *Sleeping Beauty* gene transfer, providing stable transgene expression[22]. While all of mESC colonies homogeneously express GFP (Fig E7a), only ~4% of cells in each hESC colony show a strong GFP signal (GFP(high)), indicating cellular heterogeneity (Figs. E7e, E7h-j). The fraction either weakly or unexpressing GFP we term GFP(low) and GFP(-) respectively (Fig 4a, E7b, E7e). RNAseq data of hESCs from single cells[23,24] and hPSC lines confirm that pluripotent cultures exhibit variability in HERVH expression (Fig. E1d), indicating that the GFP(high) subpopulation may differ from the GFP(low) subpopulations. Consistent with a naïve-like state, data mining of single cell RNAseq datasets[24] reveals that the expression level of HERVH in hESCs is correlated with several pluripotency-associated genes, including naïve-associated TFs (Fig. E1e).

To collect uniform GFP(high) and GFP(low) hPSCs, we performed two rounds of FACS (Fig. 4a). We first sorted GFP(+) cells that were further divided into GFP(high) and GFP(low) categories. Strikingly, GFP(high) cells are capable of forming tight, uniformly expressing 3D colonies characteristic of naïve mESCs (Fig. 4a; Supplementary Video S1). In contrast, GFP(low) cells form flat colonies, resembling mouse epiblast stem cells (mEpiSCs) (Fig. 4a). We also observed mosaic colonies. Immunostaining of 3D and chimeric colonies reveals that the NANOG and GFP(high) signals copresent (Supplementary Videos S1-2). Thus, the GFP(high) subpopulation in human pluripotent stem cells are enriched for cells resembling the murine naïve/ground state.

To examine this possibility, GFP(high) vs GFP(low) cells were subjected to expression analyses. qRT-PCR revealed significant up-regulation of naïve-associated TFs[4-6] and down-regulation of lineage-commitment genes in GFP(high) vs GFP(low) (Fig. 4b). As in naïve mESCs[25] and human ICM[26] X chromosomes are activated in GFP(high) hESCs_H9, as evidenced by nearly complete loss of condensed H3K27me3 nuclear foci (Fig. 4d) and low level of XIST expression (Fig 4c). However, nearly 60% GFP(low) hESCs transited from GFP(high) hESCs are marked with condensed H3K27me3 foci or higher density of H3K27me3 in the nucleus (Figs. 4d, E8g). These data are consistent with a naïve-like state for GFP(high) cells and a primed state for GFP(low) cells (one X chromosome inactivated or in process of being inactivated).

GFP(high) cells can be maintained in the modified 2i/LIF medium for a long time, with higher single-cell clonality as well as full pluripotency (Fig. E8a-d). However, GFP(high) and GFP(low) cells have slightly different differentiation potential. When differentiation triggered, certain naïve-associated TFs are maintained at higher levels in GFP(high) naïve-like cells compared with GFP(low), and start their differentiation program with a delay (Figs. E8e-f). Early passage hPSC cultures behave somewhat similarly to GFP(high) cells (Figs. E9a-c).

Transcriptomes of GFP-sorted cell populations and previously characterized naive-like and primed hPSCs[4] and mouse counterparts as well as human ICM, support a naive-like status of GFP(high) cells. Unbiased hierarchical clustering of the expression profiles revealed that GFP(high) and GFP(+) cells have a similar, but non-identical, expression pattern, one that sharply contrasts with GFP(low) (Fig. E8h). Strikingly, GFP(high) and GFP(+) samples clustered with human ICM and the published naïve-like hPSCs, respectively (Fig. 4e). Importantly, GFP(high) cells cluster closest to human ICM (Fig. 4e).

Cross-species comparison of expression of 9,583 mouse–human orthologs revealed that GFP(high) and GFP(+) correlated to published naïve hPSCs, while GFP(low) clustered with primed cells (Figs. 4f-g), supporting the significance of HERVH-driven transcription defining a naïve-like state.

To address how gene expression changes up to the ICM stage, we analysed 114 RNAseq samples harvested in early developmental stages of embryogenesis[24] and 3 RNAseq samples of naïve-like hESCs (3iL_hESC[3]). HERVH expression appears already in the zygote, but the pattern of activated loci changes during early development (Figs. E9d-e). Importantly, the pattern of active loci characteristic of ICM is the closest to naïve-like hESCs, including GFP(high) (Fig. E9d). Notably, the number of activated HERVH loci is particularly high in hESCs, especially in naïve-like cells and marked with H3K4me3 (Figs. E9d-f), indicating that HERVH may play some roles in the derivation and/or maintenance of naïve-like hPSCs.

To address how HERVH-driven gene expression modulates pluripotency, we surveyed differentially regulated genes in GFP(high) vs GFP(low), intersected by HERVH cis-regulation. The differentially regulated genes located in the neighbourhood (+/-50 kb) of HERVH display a similar expression pattern to those differentially expressed in GFP(high) vs GFP(low) and in human naïve-like vs primed stages, derived under specific culture conditions[4] (Fig. E9h). In contrast, a distinct pattern is observed when comparing mESCs vs mEpiSCs (Fig. E9g). Strikingly, there is an inverse pattern of expression between genes defining naïve-like stage [up in GFP(high) vs GFP(low)] and those that are down-regulated in HERVH knockdowns (rho=-0.6, P<<0.0001; Fig. E9i), underlying the significance of HERVH in regulating the naïve-like state in humans. Differentially expressed genes between GFP(high) vs GFP(low) populations were enriched for Gene Ontology (GO) terms of developmental processes, morphogenesis and organismal processes (Fig. E9j). Transition of naïve-like cells into primed state following depletion of HERVH supports the above conclusion (Fig E9k).

While GFP(high) cells have many properties resembling naïve mESCs, they are better regarded as being naïve-like, not least because it is unclear that human and naïve mESCs need be identical. Indeed, while LBP9 is associated with pluripotency[6,20] in mammals, HERVH was recruited to the pluripotency network exclusively in primates. How then to define naïve human pluripotency if we do not necessarily expect them to be identical to mouse ones? We suggest that, rather than hard to replicate inter-species chimaera experiments[27], the optimal approach is to define cells by similarity of expression to the ICM (see Supplementary Discussion). In this regard GFP(high) cells are one of the best current models of naïve-like status.

That LBP9 forms heteromer complexes functioning either as a transcriptional activator or a repressor, depending upon the partner[28] is consistent with HERVH being recruited to the pluripotency network by serendipitous modification of a pluripotency factor detailed to defend the cell against it (Fig. E10). Whatever the origin, LTR7/HERVH is an efficient reporter for the naïve-like state most probably because it acts as a platform for multiple key pluripotent transcription factors[29]. Similarly the LTR7-GFP reporter should enable optimization of naïve-like hPSC culture conditions.

Acknowledgements
LDH is Wolfson Royal Society Research Merit Award Holder. ATG is funded by a scholarship from the University of Bath. ZIz is funded by ERC-2011-AdG 294742. GGS is funded by DFG grant SCHU1014/8-1 and LOEWE Center for Cell and Gene Therapy Frankfurt/Hessian Ministry of Higher Education, Research and the Arts (ref #: III L 4-518/17.004 [2013]). We thank U. Martin and S. Merkert (Leibniz Research Laboratories for Biotechnology and Artificial Organs [LEBAO], Hannover Medical School, Hannover, Germany) for providing the cell lines hCBEC, hCBiPS1, hCBiPS2 and hiPS-SK4. Thanks to G. Klein (Karolinska Inst.) for the inspiration of working with ERVs and Z. Cseresnyés for his assistance in imaging.

Author contributions
This project was inspired by MCL. ZIz, LDH and JW conceived ideas for the project, and wrote the manuscript with contributions from other authors. The project was supervised by ZIz and LDH. ZIv provided critical advice. JW designed and performed experiments, analyzed and interpreted data, and participated in bioinformatic analyses. TR contributed by EMSA and assisted in immunostaining experiments. AS assisted in the reporter assays. HC assisted in shRNA cloning. WC and JW performed RNAseq experiments. AP provided materials, and performed karyotype analysis. DB, NVF and GGS provided materials. GX performed RNAseq, bisulfite-seq and ChIP-seq analyses. MS analyzed microarray data and performed cross-species correlation studies. LDH and ATG performed all the other bioinformatic analyses.

RNASeq and microarray data were submitted to NCBI's GEO database at accession GSE54726.

Ethics approval
For work on human ES cells we obtained No.6 allowance from the Robert Koch Institute, Germany (08.10.2004). The human embryonic stem cell lines (H1, H9, BGN1, and BGN2) are permitted to be used in the study "Mechanisms of single transduction in the maintenance of undifferentiated state in human embryonic stem cells."

The author's declare no competing financial interests.

References

1 Welling, M. & Geijsen, N. Uncovering the true identity of naive pluripotent stem cells. *Trends Cell Biol.* **23**, 442-448, doi:10.1016/j.tcb.2013.04.004 (2013).

2 Ware, C. B. *et al.* Derivation of naïve human embryonic stem cells. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1319738111 (2014).

3 Chan, Y. S. *et al.* Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13**, 663-675, doi:10.1016/j.stem.2013.11.015 (2013).

4 Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282-286, doi:10.1038/nature12745 (2013).

5 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci USA* **107**, 9222-9227, doi:10.1073/pnas.1004584107 (2010).

6 Martello, G., Bertone, P. & Smith, A. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. *The EMBO journal*, doi:10.1038/emboj.2013.177 (2013).

7 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631-634, doi:10.1038/ng.600 (2010).

8 Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423-425, doi:10.1038/nsmb.2799 (2014).

9 Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558-566, doi:10.1038/ng.2965 (2014).

10 Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57-63, doi:10.1038/nature11244 (2012).

11 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, doi:10.1186/1742-4690-9-111 (2012).

12 Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**, R107, doi:10.1186/gb-2012-13-11-r107 (2012).

13 Gaspar-Maia, A. *et al.* Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-868, doi:10.1038/nature08212 (2009).

14 Chappell, J., Sun, Y., Singh, A. & Dalton, S. MYC/MAX control ERK signaling and pluripotency by regulation of dual-specificity phosphatases 2 and 7. *Genes Dev.* **27**, 725-733, doi:10.1101/gad.211300.112 (2013).

15 Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-1148, doi:10.1016/j.cell.2013.04.022 (2013).

16 van den Berg, D. L. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6**, 369-381, doi:10.1016/j.stem.2010.02.014 (2010).

17 Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**, 1106-1117, doi:http://dx.doi.org/10.1016/j.cell.2008.04.043 (2008).

18 Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113-1117, doi:10.1038/ng.710 (2010).

19 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. *The EMBO journal* **31**, 522-533, doi:10.1038/emboj.2011.459 (2012).

20 Ye, S., Li, P., Tong, C. & Ying, Q. L. Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. *The EMBO journal*, doi:10.1038/emboj.2013.175 (2013).

21 Nishiyama, A. *et al.* Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Scientific reports* **3**, 1390, doi:10.1038/srep01390 (2013).

22 Mates, L. *et al.* Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41**, 753-761, doi:10.1038/ng.343 (2009).

23 Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-782, doi:10.1038/nbt.2282 (2012).

24 Yan, L. Y. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131-+, doi:10.1038/nsmb.2660 (2013).

25 Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487-492, doi:10.1016/j.stem.2009.05.015 (2009).

26 Okamoto, I. *et al.* Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* **472**, 370-374, doi:10.1038/nature09872 (2011).

27 Theunissen, Thorold W. *et al.* Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. *Cell Stem Cell*, doi:10.1016/j.stem.2014.07.002 (2014).

28 To, S., Rodda, S. J., Rathjen, P. D. & Keough, R. A. Modulation of CP2 family transcriptional activity by CRTR-1 and sumoylation. *PloS one* **5**, e11702, doi:10.1371/journal.pone.0011702 (2010).

29 Dunn, S. J., Martello, G., Yordanov, B., Emmott, S. & Smith, A. G. Defining an essential transcription factor program for naive pluripotency. *Science* **344**, 1156-1160, doi:10.1126/science.1248882 (2014).

For online methods and Supplementary Data

30 Grabundzija, I. *et al.* Sleeping Beauty transposon-based system for cellular reprogramming and targeted gene insertion in induced pluripotent stem cells. *Nucleic Acids Res*, doi:10.1093/nar/gks1305 (2012).

31 Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8**, 444-445, doi:10.1038/nmeth.1611 (2011).

32 Hanna, J. *et al.* Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell* **4**, 513-524, doi:10.1016/j.stem.2009.04.015 (2009).

33 Zhou, W. *et al.* Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. *Mol. Cell. Biol.* **20**, 7662-7672 (2000).

34 Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068-1081, doi:10.1016/j.cell.2012.08.011 (2012).

35 Haase, A. *et al.* Generation of induced pluripotent stem cells from human cord blood. *Cell Stem Cell* **5**, 434-441, doi:10.1016/j.stem.2009.08.021 (2009).

36 Prigione, A., Fauler, B., Lurz, R., Lehrach, H. & Adjaye, J. The senescence-related mitochondrial/oxidative stress pathway is repressed in human induced pluripotent stem cells. *Stem Cells* **28**, 721-733, doi:10.1002/stem.404 (2010).

37 Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872, doi:10.1016/j.cell.2007.11.019 (2007).

38 Onder, T. T. *et al.* Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483**, 598-602, doi:10.1038/nature10953 (2012).

39 Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-510 (1997).

40 Kaufman, C. D., Izsvak, Z., Katzer, A. & Ivics, Z. Frog Prince transposon-based RNAi vectors mediate efficient gene knockdown in human cells. *J RNAi Gene Silencing* **1**, 97-104 (2005).

41 Wang, Z., Oron, E., Nelson, B., Razis, S. & Ivanova, N. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell* **10**, 440-454, doi:10.1016/j.stem.2012.02.016 (2012).

42 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

43 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

44 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

45 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

46 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

47 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

48 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:10.1038/nbt.1662 (2010).

49 Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618-1619, doi:10.1093/bioinformatics/btu056 (2014).

50 Frith, M. C. *et al.* Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32**, 1372-1381, doi:10.1093/nar/gkh299 (2004).

51 Haverty, P. M., Hansen, U. & Weng, Z. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res* **32**, 179-188, doi:10.1093/nar/gkh183 (2004).

52 Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90, doi:10.1038/nature11212 (2012).

53 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

54 Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, doi:10.1093/nar/gks915 (2012).

55 Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345-W349, doi:10.1093/nar/gkm391 (2007).

56 Vassena, R. *et al.* Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699-3709, doi:10.1242/dev.064741 (2011).

Figure 1. HERVH is a specific marker of human pluripotent stem cells (hPSCs)

a, Expression of various Transposable Elements (TEs) in human induced pluripotent stem cells (hiPSC), hESC (H1), and human fibroblast HFF-1. Colours indicate different classes of TEs (red, long terminal repeat elements (LTR); green, long interspersed nuclear elements (LINE); blue, short interspersed nuclear elements (SINE); grey, other repeat elements). **b,** The proportion of active loci in each HERV family. **c,** Relative mRNA levels of HERV(H/K/W) in hESC (HES-3), various hiPSCs lines and their parental somatic cells. **d,** Effect of long-term culturing on HERVH transcription levels in hiPSCs generated from HFF-1. P, passage number. **c**, **d**, mRNA levels are normalized to GAPDH, and relative to HES-3. Error bars, s.d. (n=3 independent cell cultures), t-test, *P<0.05.

Figure 2. HERVH is recruited into the circuitry of human pluripotency

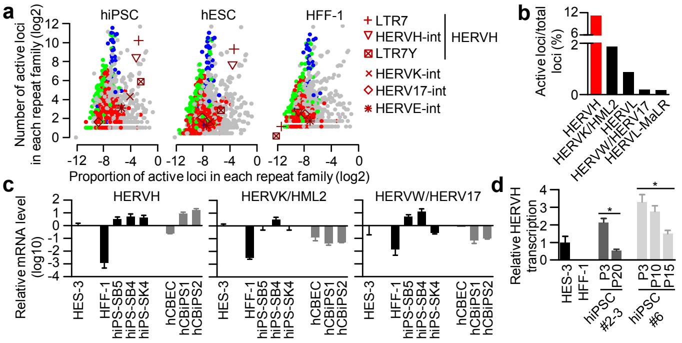
a, The distribution of H3K4me3 and H3K9m3 in active vs inactive HERVH regions in hiPSCs, hESCs and HFF-1. **b,** Conserved binding sites of OCT4, NANOG, LBP9 and KLF4 are shown in active LTR7s vs moderately active versions of LTR7Y/C. The Jaspar consensus sequence of the LBP9 is shown. **c,** Confirmation of LBP9 binding to LTR7 by ChIP-qPCR with two different primers (LTR7#1, #2) targeting LTR7 regions. HERVH-gag, HERVH-pol and LTR5_Hs (LTR of HERVK) served as negative controls, while an upstream region of NANOG (7.5 kb from TSS) was a positive control. Data are collected from two independent experiments with biological replicates per experiment (LBP9: n=3; IgG: n=2), error bars, s.d.; t-test *P<0.05, **P<0.01. **d,** Upregulation of HERVH transcription in HFF-1 regulated by exogenous pluripotency-associated transcription factors. Data are collected from three independent experiments with biological triplicates per experiment. **e-f,** Effects of shRNA knockdowns of various TFs on HERVH and HERVK transcription in hESC_H9. Data shown are representative of three independent experiments with biological triplicates per experiment. **d**-**f**, error bars, s.d.; t-test *P<0.05, **P<0.01, P***<0.001.

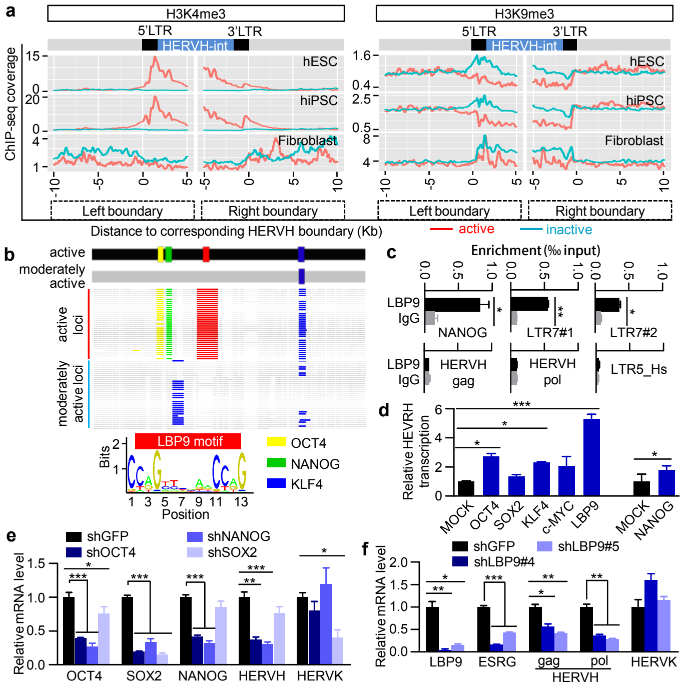
Figure 3. HERVH triggers pluripotency-regulating hPSC-specific chimeric transcripts and lncRNAs

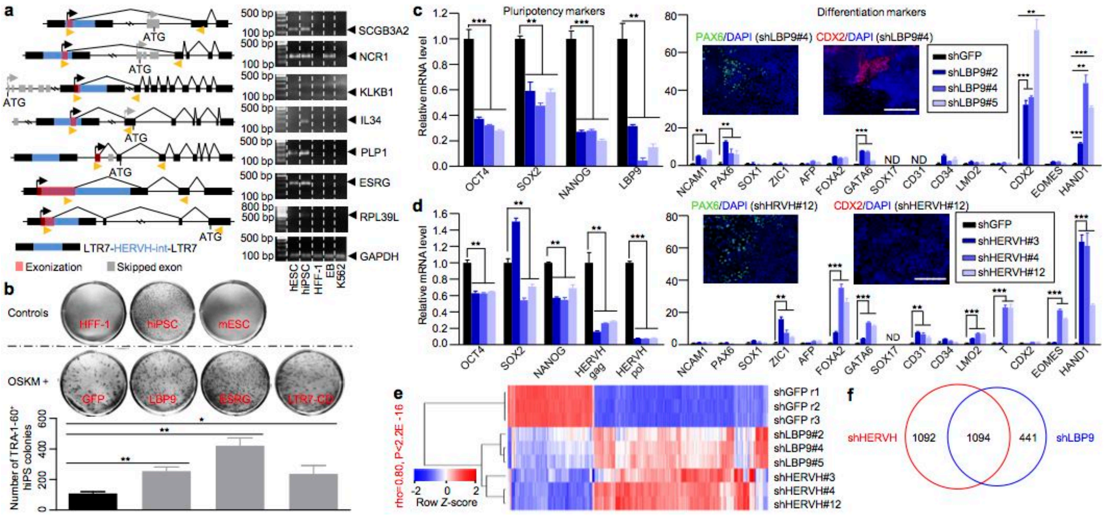
a, Expression of HERVH forces diversification of transcripts in hPSCs. Left: schematic representation of the HERVH-derived alternative and chimeric transcripts. Right: RT-PCR detects HERVH-specific transcripts (marked by triangles) in hPSCs and NCR1 in embryoid body (EB), but not in HFF-1 or K562. Yellow arrows indicate primer binding sites. **b,** The effects of LBP9 and HERVH-derived transcripts on reprogramming of HFF-1 to hiPSCs. Upper panel: Representative TRA-1-60 stained wells are shown. Lower panel: The number of TRA-1-60$^+$ hiPS colonies reprogrammed from HFF-1 by LBP9, ESRG or LTR7-CD in conjunction with OCT4, SOX2, KLF4 and c-MYC (OSKM). Error bars, s.d., t-test *P<0.05, **P<0.01 from three independent experiments. **c-d,** qRT-PCR analyses to determine the relative expression level of pluripotency and differentiation markers after knockdown of LBP9 (**c**) or HERVH (**d**) in hESC_H9. Data shown are representative of three independent experiments with biological triplicates per experiment. Error bars, s.d., t-test *P<0.05, **P<0.01, and ***P<0.001. ND, not detected. Representative immunostainings show the expression of PAX6 and CDX2 in LBP9 and HERVH knockdowns (scale bar, 100 μm). **e,** Heat map showing genome-wide gene expression in hESC_H9 following knockdown of GFP (shGFP), LBP9 (shLBP9) and HERVH (shHERVH). The knockdown effect of LBP9 and HERVH are highly similar (rho from Spearman's correlation). For list of affected genes, including direct targets of shHERVH see Tables S13 and S14. **f,** Venn diagram shows that 1094/2627 genes are similarly affected by KD-HERVH and KD-LBP9 (Table S12).

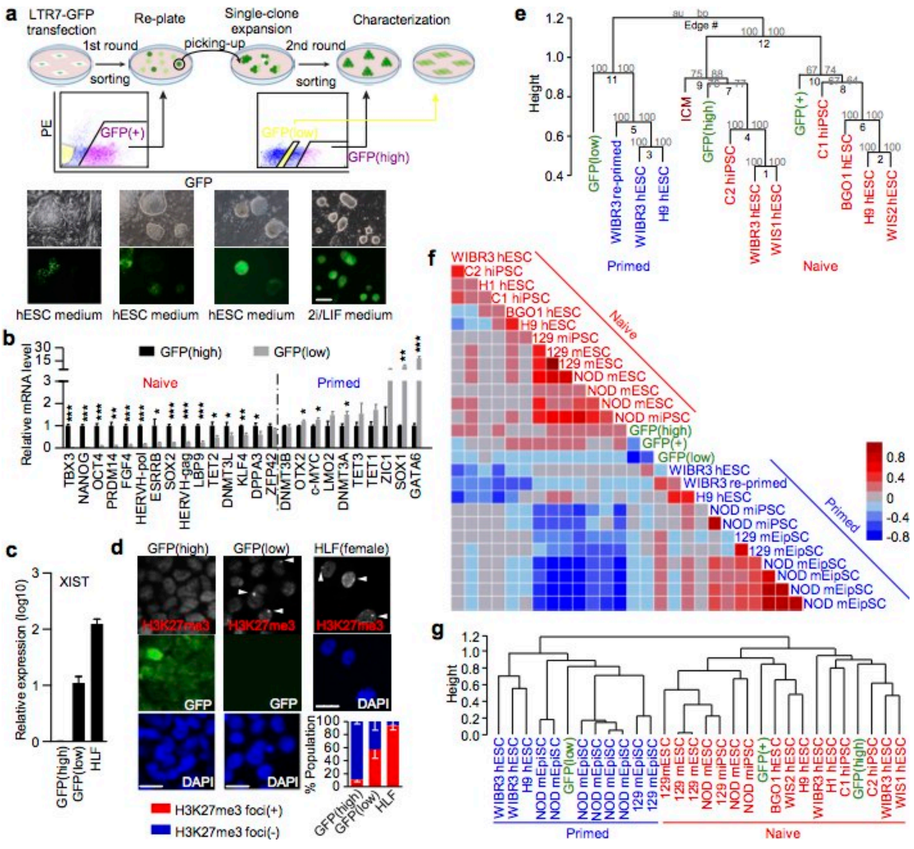
Figure 4. HERVH genetically marks naïve-like hESCs

a, Experimental scheme for isolating naïve-like hPSCs. pT2-LTR7-GFP#2-marked hESC_H9 were enriched by FACS-sorting in multiple rounds and cultured in *conventional hESC medium* and in 2i/LIF medium, respectively. Scale bar, 200 *μ*m. See also Supplementary Videos S1 and S2. **b,** qRT-PCR analyses of multiple transcription factors and markers for naive and primed state in GFP(high) and GFP(low) cells, respectively. **c,** qRT-PCR analysis of XIST in GFP(high), GFP(low) hESC_H9 and human female fibroblasts (HLF). **b**, **c**, Error bars, s.d.; t-test *P<0.05, **P<0.01, and ***P<0.001(n=3 independent cell cultures). **d**, Representative confocal images obtained after immunostaining for H3K27me3 on GFP(high), GFP(low), hESC_H9s and HLF. Scale bar, 20 μm. The proportions of H3K27me3 foci(+) (triangles) and (-) cells in each sample are shown in the histogram. Error bar, s.d.. Data were obtained from 100-450 cells counted from five images per sample. **e**, Global expression cluster dendrogram between GFP(high), GFP(+), GFP(low) hESCs_H9, human inner cell mass (ICM) and previously established human naïve and primed cell lines[4]. Approximately Unbiased (AU) probability, Bootstrap Probability (BP) values and edge numbers at P-value less than 0.01 are shown. ICM clusters closest with GFP(high) – nodes 7,9. **f,** Correlation matrix displaying the unbiased and pairwise comparison of mouse–human orthologous gene expression between GFP-marked hESC_H9 (this study, green) and mouse and human[4] naïve as well as primed PSCs. Color bar indicates Spearman correlation strength. **g,** Cluster analysis using the average distance method on the same dataset as in **f**. GFP(high), GFP(+) and GFP(low) cells in **e-g** were collected from hESC_H9 cells cultured in *conventional human ESC medium* by FACS-sorting.


a
hiPSC
hESC
HFF-1
Number of active loci in each repeat family (log2)
0
2
4
6
8
10
12
-12
-8
-4
0
Proportion of active loci in each repeat family (log2)
+ LTR7
▽ HERVH-int
⊠ LTR7Y
HERVH
× HERVK-int
◇ HERV17-int
✱ HERVE-int
b
Active loci/total loci (%)
0
2
6
12
HERVH
HERVK/HML2
HERVL
HERVW/HERV17
HERVL-MaLR
c
HERVH
HERVK/HML2
HERVW/HERV17
Relative mRNA level (log10)
2
1
0
-1
-2
-3
-4
HES-3
HFF-1
hiPS-SB5
hiPS-SB4
hiPS-SK4
hCBEC
hCBiPS1
hCBiPS2
d
Relative HERVH transcription
0
1
2
3
4
*
HES-3
HFF-1
P3
P20
hiPSC #2-3
P3
P10
P15
hiPSC #6


a
H3K4me3
H3K9me3
5'LTR
3'LTR
HERVH-int
ChIP-seq coverage
hESC
hiPSC
Fibroblast
Left boundary
Right boundary
Distance to corresponding HERVH boundary (Kb)
active
inactive
b
active
moderately active
active loci
moderately active loci
LBP9 motif
Bits
Position
OCT4
NANOG
KLF4
c
Enrichment (‰ input)
LBP9
IgG
NANOG
LTR7#1
LTR7#2
HERVH gag
HERVH pol
LTR5_Hs
d
Relative HEVRH transcription
MOCK
OCT4
SOX2
KLF4
c-MYC
LBP9
MOCK
NANOG
e
shGFP
shOCT4
shNANOG
shSOX2
Relative mRNA level
OCT4
SOX2
NANOG
HERVH
HERVK
f
shGFP
shLBP9#4
shLBP9#5
Relative mRNA level
LBP9
ESRG
gag
pol
HERVH
HERVK


a
ATG
500 bp
100 bp
800 bp
SCGB3A2
NCR1
KLKB1
IL34
PLP1
ESRG
RPL39L
GAPDH
hESC
hiPSC
HFF-1
EB
K562
LTR7-HERVH-int-LTR7
Exonization
Skipped exon
b
Controls
HFF-1
hiPSC
mESC
OSKM +
GFP
LBP9
ESRG
LTR7-CD
Number of TRA-1-60+ hiPS colonies
0
200
400
600
*
**
c
Pluripotency markers
Differentiation markers
Relative mRNA level
OCT4
SOX2
NANOG
LBP9
PAX6/DAPI (shLBP9#4)
CDX2/DAPI (shLBP9#4)
shGFP
shLBP9#2
shLBP9#4
shLBP9#5
NCAM1
PAX6
SOX1
ZIC1
AFP
FOXA2
GATA6
SOX17
CD31
CD34
LMO2
T
CDX2
EOMES
HAND1
ND
***
d
HERVH gag
HERVH pol
PAX6/DAPI (shHRVH#12)
CDX2/DAPI (shHERVH#12)
shHERVH#3
shHERVH#4
shHERVH#12
e
rho=0.80, P<2.2E -16
-2
0
2
Row Z-score
shGFP r1
shGFP r2
shGFP r3
shLBP9#2
shLBP9#4
shLBP9#5
f
shHERVH
1092
1094
441
shLBP9


a
LTR7-GFP transfection
Re-plate
Single-clone expansion
Characterization
1st round
sorting
picking-up
2nd round
sorting
PE
GFP(+)
GFP(low)
GFP(high)
GFP
hESC medium
hESC medium
hESC medium
2i/LIF medium
b
Relative mRNA level
GFP(high)
GFP(low)
Naive
Primed
TBX3
NANOG
OCT4
PRDM14
FGF4
HERVH-pol
ESRRB
SOX2
HERVH-gag
LBP9
TET2
DNMT3L
KLF4
DPPA3
ZFP42
DNMT3B
OTX2
c-MYC
LMO2
DNMT3A
TET3
TET1
ZIC1
SOX1
GATA6
c
XIST
Relative expression (log10)
GFP(high)
GFP(low)
HLF
d
GFP(high)
GFP(low)
HLF(female)
H3K27me3
GFP
DAPI
% Population
H3K27me3 foci(+)
H3K27me3 foci(-)
e
Height
au
bp
Edge #
Primed
Naive
GFP(low)
WIBR3 re-primed
WIBR3 hESC
H9 hESC
ICM
GFP(high)
C2 hiPSC
WIBR3 hESC
WIS1 hESC
GFP(+)
C1 hiPSC
BGO1 hESC
H9 hESC
WIS2 hESC
f
Naive
Primed
WIBR3 hESC
C2 hiPSC
H1 hESC
C1 hiPSC
BGO1 hESC
H9 hESC
129 miPSC
129 mESC
129 mESC
NOD mESC
NOD mESC
NOD mESC
NOD miPSC
GFP(high)
GFP(+)
GFP(low)
WIBR3 hESC
WIBR3 re-primed
H9 hESC
NOD miPSC
NOD miPSC
129 mEipSC
129 mEipSC
NOD mEipSC
NOD mEipSC
NOD mEipSC
NOD mEipSC
g
Height
WIBR3 hESC
WIBR3 hESC
H9 hESC
NOD mEpiSC
NOD mEpiSC
GFP(low)
NOD mEpiSC
NOD mEpiSC
NOD mEpiSC
NOD mEpiSC
129 mEpiSC
129 mEpiSC
129mESC
129 mESC
129 mESC
NOD mESC
129 miPSC
NOD mESC
NOD miPSC
GFP(+)
BGO1 hESC
WIS2 hESC
H9 hESC
WIBR3 hESC
H1 hESC
C1 hiPSC
GFP(high)
C2 hiPSC
WIBR3 hESC
WIS1 hESC
Primed
Naive