The phylogenetically distinct early human embryo

Manvendra Singh[1], Thomas J. Widmann[2], Jose L. Cortes[2], Gerald G. Schumann[3], Stephanie Wunderlich[4], Ulrich Martin[4], Jose L. Garcia-Perez[2,5], Laurence D. Hurst[6*], Zsuzsanna Izsvák[1*]

[1] Max-Delbrück-Center for Molecular Medicine in the Helmholtz Society, Robert-Rössle-Strasse 10, 13125 Berlin, Germany.
[2] GENYO. Centre for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Regional Government, PTS Granada, 18016 Granada, Spain.
[3] Paul-Ehrlich-Institute, Division of Medical Biotechnology, Paul-Ehrlich-Strasse 51-59, 63225 Langen, Germany.
[4]Center for Regenerative Medicine Hannover Medical School (MHH) Carl-Neuberg-Str.1, Building J11, D-30625 Hannover, Germany
[5]Institute of Genetics and Molecular Medicine (IGMM), University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, United Kingdom
[6] The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, Somerset, UK, BA2 7AY.

*Corresponding authors

*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

The Milner Centre for Evolution,
Department of Biology and Biochemistry,
University of Bath
Bath, Somerset, UK BA2 7AY
tel: +44 (0)1225 386424
Fax: +44 (0)1225 386779
email: l.d.hurst@bath.ac.uk
> http://people.bath.ac.uk/bssldh/LaurenceDHurst/Home.html

Email addresses:
Singh, Manvendra < Manvendra.Singh@mdc-berlin.de
Widmann, Thomas < thomas.widmann@genyo.es
Schumann, Gerald < Gerald.Schumann@pei.de
Cortes, Jose L., <jlcortesromero@gmail.com>
Wunderlich.Stephanie@mh-hannover.de <Wunderlich.Stephanie@mh-hannover.de>
Martin,Ulrich@mh-hannover.de < Martin.Ulrich@mh-hannover.de
Garcia-Perez, Jose <Jose.Garcia-Perez@igmm.ed.ac.uk>
Hurst, Laurence < L.D.Hurst@bath.ac.uk
Izsvak, Zsuzsanna < zizsvak@mdc-berlin.de

- The human blastocyst has an inner cell mass, despite claims to the contrary
- Cell purging via apoptosis defines a phylogenetically restricted class of blastocyst non-committed cells

- Fast transcriptome evolution is relatively unique to the pluripotent epiblast and is mostly due to a primate-specific transposable element
- Current naive cultures don't reflect human uniqueness, are heterogeneous and are developmentally "confused".

Abstract

The phylogenetic singularity of the human embryo remains unresolved as cell types of the human blastocyst have resisted classification. Combining clustering of whole cell transcriptomes and differentially expressed genes we resolve the cell types. This unveils the missing inner cell mass (ICM) and reveals classical step-wise development. Conversely, numerous features render our blastocyst phylogenetically distinct: unlike mice, our epiblast is self-renewing and we have blastocyst non-committed cells (NCCs), part of an apoptosis-mediated quality control/purging process. At the transcriptome-level all primate embryos are distinct as the pluripotent cell types are uniquely fast evolving. All major gene expression gain and loss events between human and new-world monkeys involve endogenous retrovirus H (ERVH). Human pluripotent cells are unique in which (H)ERVH's are active, the extent to which these modulate neighbour gene expression and their ability to suppress mutagenic transposable elements. Current naïve cultures are heterogeneous and both developmentally and phylogenetically "confused".

Introduction

Blastocysts harbour the blueprints of the body plan as they have the potential to give rise to all somatic and germ cell lineages (Gardner, 1998). The most potent cells are thought to be the inner cell mass (ICM) that gives rise to the hypoblast (primitive endoderm) and the pluripotent epiblast (EPI) (Gardner, 1998). In contrast to mice (Flach et al., 1982), the primate embryonic gene activation (EGA) occurs not in the zygote, but later (4/8 cells in human) (Braude et al., 1988). Early human development is assumed to be unusual in that an ICM is not evidenced, leading to speculation that such a mass of cells may not exist as a distinct lineage (Petropoulos et al., 2016; Sahakyan and Plath, 2016). In addition, the segregation of morula into TE, EPI, and primitive endoderm (PE) thought to occur simultaneously rather than the step-wise manner seen in mice or macaques (Chazaud and Yamanaka, 2016; Niakan et al., 2012; Sahakyan and Plath, 2016).

While it has been a consensus that human pre-implantation embryogenesis is exceptional (Nakamura et al., 2016),

even using single cell data (Blakeley et al., 2015; Petropoulos et al., 2016; Yan et al., 2013), the cells of the blastocyst, however, still resist unambiguous identification. Owing to the fast progression of the developmental process, suddenly generating a large number of cells, the most challenging task is to catch the distinct lineages during blastocyst formation. To address this, and to enable characterization of all stages of early human development, we use a strategy of clustering both cells and genes. We employ a dimension-reduction clustering methodology on single cell transcriptomics (Satija et al., 2015) that focuses on defining local resemblances. This strategy is ideally suited to single cell data, to resolving transitory stages, capturing unresolvable clusters and, in turn, to identifying diagnostic markers, hence informing the biology of discrete cell types. Expecting a considerable turnover in gene expression during primate pre-implantation embryogenesis, we also employ comparative single cell high-resolution analysis of human vs macaque (*Cynomolgus fasciculara*) transcriptomes. Using the above approaches, we ask if the cell types are preserved and whether all the cell types of early embryogenesis are equally subject to evolutionary processes. Our strategy (i) unambiguously detects the human ICM, (ii) identifies cells expressing multiple lineage markers, (iii) resolves a novel non-committed cell type during the formation of the blastocyst that filtered out from the developmental process and (iv) reveals that while the pluripotent ICM/EPI (mostly EPI), is highly divergent between primates, the primitive endoderm (PE) and TE remain almost unchanged. To decipher the mechanism of accelerated evolution of EPI, we take the opportunity that the EPI can be modelled using cultured pluripotent cells (PSCs). We follow the evolution of pluripotency by examining gene expression and structural differences in more detail between old-world (OWMs) and new world monkeys (NWMs) in selected primate PSCs, including human, bonobo, gorilla and marmoset (*Callithrix*).

The OWMs share a high degree of similarity with humans in their genome sequence (92.5% to 97.5%) (Olson and Varki, 2003; Yan et al., 2013). The presence/absence differences are in no small part attributed to transposable element (TrE) insertions (Ramsay et al., 2017), including those derived from multiple waves of retroviral invasions into primate genomes (endogenous retroviruses, ERVs). After EGA, reactivating the transcription of TrE families of different phylogenetic age is evidenced, and has a characteristic patterning (Friedli and Trono, 2015; Goke et al., 2015; Grow et al., 2015; Guo et al., 2014; Izsvak et al., 2016; Rowe and Trono, 2011; Smith et al., 2014). To decipher if (and how) TrEs contributed the primate evolution of the pre-implantation development, in addition to considering classical genes as markers, we analyse the transcriptional dynamics of transcripts associated with TrEs. We find that both the young transpositionally active (e.g. genotoxicity in NCC) and older transpositionally inactive (e.g. rapid evolution of the ICM/EPI) TrE-derived transcriptional changes are associated with the primate evolution of the blastocyst.

Finally, a new classification of early cell types and markers enables us to ask whether there is a pluripotent cell type in the human blastocyst that would be a good candidate for extraction and stable maintenance *in vitro* or to mimic. In characterizing recently derived naïve cultures we find them developmentally and phylogenetically confused.

Results

Resolving the identity of cells expressing multiple and no lineage markers unmasks the human inner cell mass (ICM)

What cell-types are present in the human pre-implantation embryo? In particular, is the inner cell mass really missing from human blastocysts? To classify cell-types we used available single cell transcriptome data (Petropoulos et al., 2016; Yan et al., 2013). In contrast to these previous analyses that identified differentially expressed genes (DEGs), we used a strategy of clustering both whole cell trancriptomes and DEGs, using a combination of clustering K-means and principal component (PCs). In doing that, we identified 1597 genes exhibiting high variability across single cells and thus potentially useful for defining cell types (Figures 1A and S1A). Next, we performed principle component analysis (PCA) to reduce the dimensionality of the data and identified nine significant principal components (PC) using a previously described 'jackstraw' method (Satija et al., 2015). We used these PC loadings as inputs to t-distributed stochastic neighbour embedding (t-SNE) for visualization. This approach allowed us to distinguish 10 clusters that we annotate on the basis of previously reported expressed markers (Petropoulos et al., 2016) (Figures 1A and S1A).

In E3-E4, it is relatively straightforward to identify the clusters such as oocytes, zygote, 2, 4, 8 cells stage (E3) and even the more heterogeneous morula (E4) of two subgroups (Figure S1B). While LEUTX1 flags the 8-cell stage, the two clusters of morula are marked by either HKDC1 or GATA3 (Figures S1C), the latter is further traceable in Pre-TE. Human blastocyst formation initiates at E5, progresses at E6 and stabilizes at E7 prior to implantation (Figures S1D). After morula our analysis reveals previously unidentified clusters (Figures 1A). Our strategy distinctly identifies EPI and PE, as well as TE (polar, mural) clusters in E5, E6 and in E7 stages, respectively [area under curve (AUC) ≥ 0.90] (Figures 1A and S1E). A remaining cluster from E6 and one from E6-E7 is yet to be defined since they express markers heterogeneously (Figures 1A and S1A).

Not all the clusters are so easy to classify as EPI, PE and the early cell types. Given heterogeneity in a cluster, might some cells also be transitory types? At E5, we observed two types of cells, either expressing (even multiple) lineage markers or those that fail to express any markers of known blastocyst lineages (EPI-PE-TE). Cells expressing multiple markers at E5 do not segregate on the t-SNE plot using unbiased approaches of clustering E3-E7 cells, and we defined them as transitory cells (Figure 1A). By contrast, cells expressing none of the known markers form a clearly segregated cell population on the t-SNE plot that we name non-committed cells (NCCs) (Figures 1A and S1A).

If ICM exists, we would expect it to be resolved as a cell type segregating from morula. In order to further resolve the broad spectrum of transcriptomes of cells segregating from morula, we restricted analysis to E5 only, and removed cells with low quality transcriptomes (expressing (log2 Transcript Per Million (TPM > 1) less than 5000 genes). The transcriptomes of the remaining 300 cells were subjected to a similar strategy that we used for dissecting E3-E7. This approach resulted in six significant principal components (PCs), and we enlisted the top 30 genes contributing to their respective eigen vectors (Figure S2C). Loading the above PCs as input, we observe three distinct transcriptome clusters on t-SNE (Figure 1B) (that could also be distinguished on the first two principal components) using the expression dynamics of Most Variable Genes (MVGs) (Figure S2A). Altogether, we identified 235 genes (AUC > 0.80) that we used to characterize the individual clusters (Figure 1D). Expression of DLX3, a known marker of a precursor population of TE defines the first cluster (n=86) as pre-TE. Curiously, the second cluster (n=97), corresponding to the freshly identified NCC, homogeneously expresses BIK. The third cluster (n=71) co-expresses known EPI (e.g. NANOG) and PE (e.g. BMP2) markers, defining the ICM (Figures 1C-E). Thus, our strategy to resolve distinct cell populations segregating from morula enabled us to unmask the human ICM. In sum, based on their ranking in the corresponding clusters, we identify the top markers of human ICM (e.g. IL6R, SPIC), pre-TE (e.g. DLX3, TMPRSS2) and NCC populations (e.g. BIK, BAK1) (Figures 1D-E).

Human pre-implantation embryogenesis is a sequential process segregating from morula

Identifying the human ICM would challenge the recent 'simultaneous model' for human blastocyst formation. This model suggests that the human morula segregates to EPI, PE and TE simultaneously around E5 (Petropoulos et al., 2016), a deviation from the step-wise lineage specification dynamics of mouse or macaque (Chazaud and Yamanaka, 2016; Nakamura et al., 2016). To evaluate the 'simultaneous' vs 'sequential' models, we employed scaled expression of our cluster-specific markers, DLX3 (Pre-TE), BMP2 (PE), NANOG (EPI), IL6R (ICM) and BIK (NCC), and determined co-expression patterns at single cell resolution (Figures 1E).

This strategy helped us to identity of the transitory cells and decipher developmental path following morula. Out of 300 cells, we detected 3 expressing all four markers, indicating that these cells could be the precursors of early blastocysts. Another subset of cells expressed the markers of the three layers of the blastocyst (n=25), featuring a transitory/precursor state (T1) prior the segregation to ICM (n=71) and pre-TE (n=86). ICM and either PE or EPI markers were enriched exclusively in the transitory cell population of T2 (n=8) and T3 (n=10), respectively (Figures 1E and S2F). We hypothesized that these ICM-derived cells are the Pre-EPI and Pre-PE cells that would commit to PE and EPI at E6-7 stages. The identification of ICM and the transitory cells supports the model that the human early embryogenesis is a step-wise process and thus resembling that seen in mouse and macaque.

How does the transcriptome of our freshly identified human ICM compare with that of a non-human primate (NHP)? We compared single cell transcriptomes of the blastocysts from human and a macaque *(Cynomolgus fascicularis)* (Nakamura et al., 2016; Petropoulos et al., 2016). In contrast to the human study, the lineage specific cells (e.g. ICM, EPI, PE/hypoblast and TE) were extracted prior to sequencing in *Cynomolgus*, thus no NCCs were isolated. We reclassified their transcriptomes by PCA revealing a similar pattern of distinct cell types in both macaque and human (Figures S3A-B). For comparison, we only use genes annotated in both species. Using the top transcription markers that are expressed in both species, we could identify EPI (e.g. NODAL, GDF3, PRDM14), PE/hypoblast (e.g. APOA1, GATA4 and COL4A1) and TE (e.g. DLX3, STS and PGF). Using this strategy, we could identify again the ICM, unambiguously marked by the expression of SPIC in both species (Figure S3C).

Are non-committed cells parts of a quality control mechanism filtering out damaged cells?

To discern more of the biology of previously unreported non-committed cells (NCCs), we determine the markers defining them. The top marker of NCCs, BIK (BCL2-Interacting Killer) (Figures 1C-D and S2B-E), is an apoptosis-inducing factor, suggesting that this cell population have no developmental future. To clarify this, we averaged the expression for individual genes across the cell types and performed pairwise analysis, enabling identification of the genes that are differentially regulated in committed ICM vs non-committed cells. KEGG pathway mapping revealed that NCC enriched genes belong to the *Apoptosis pathway* (Figures S2B-E). In addition to BIK1, we observed the differential upregulation of numerous genes associated with programmed cell death (e.g. BAK1, various caspases or MAPK3, etc.) (Figures S2B and S2E). By contrast, ICM genes (e.g. BMP2, NANOG, etc.) were, as expected, enriched in *Pluripotency regulating signalling* (Figure 1C-E and S2B-D).

Apoptosis is thought to play an active role throughout the developmental process, although not before embryonic gene activation (EGA) (Hardy, 1999). To find out, where NCCs come from, we performed a transcriptome-wide clustering with 1000 bootstraps. Since, the cells are in continuous progression exhibiting a dynamic transcriptome in the otherwise clearly definable clusters, we averaged their expression for the clusters. This analysis detects NCCs post-morula as an alternative population of cells that fail to commit either to ICM or Pre-TE (Figure 2A).

Why might the NCCs be expressing apoptotic factors? This may be a mechanism that serves as a quality control measure to eradicate damaged cells or that simply removes unnecessary cells during the developmental process. In the former context, one possibility is that they might be damaged by the activity of mutagenic transposable elements (TrEs). In humans, the phylogenetically young elements include certain transpositionally active TrEs, such as Long Interspersed Element class 1 (LINE-1 or L1), SVA and Alu (Hancks and Kazazian, 2012; Mills et al., 2007). The majority of the young elements in the human genome are activated following EGA with their expression peaking in morula (Goke et al., 2015; Romer et al., 2017). Thus, we would expect that activation of TrEs would adversely affect some cells in the embryo.

The quality control hypothesis predicts that NCCs should express young TrEs with transposing potential, while committed cells would not. To monitor the dynamics of TrEs expression following morula, in the blastocyst, we averaged the expression (Log2 CPM+1) of each particular TrE family and compared their expression in NCCs against ICM. We detected transcriptional upregulation of TrEs in both NCC and ICM (Figure 2B). However, while the activated families in the ICM are phylogenetically old and transpositionally inactive, the upregulated TrEs in NCCs are young and include transpositionally-competent elements (Figure 2B). The list of activated young TrEs in NCCs includes both LTR-containing TrE such as LTR5_Hs and recently mutagenic non-LTR retrotransposons, such as AluY (Ya5), SVA-D/E and L1_Hs elements. To corroborate these findings, we analyzed human pre-implantation embryos using confocal microscopy and an antibody raised against the L1_Hs-encoded L1_Hs-ORF1p protein. Our immunostaining detects robust expression of the L1_Hs-ORF1p during blastocyst formation, *in situ* (Figures S3D and Movie 1) with an inverse correlation with POUF5F1 (OCT4) levels (Figure 2C). The cells are compacting to form ICM show high intensity of POU5F1 staining whereas, L1_HS-ORF1p stains scattered cells, belonging neither the forming ICM nor trophectoderm (Figures 2C and S3D), suggesting that L1_HS-ORF1p^{high} cells fail to express a commitment marker. We also detect cells with signs of genomic DNA damage, visualised by H2AXγ staining (Figure 2D). While, massive transcriptional upregulation of certain TrEs might already trigger apoptosis,

we speculate that the DNA damage in NCCs, generated by the endonuclease activity of L1_Hs contributes to the apoptotic process. The quality control hypothesis suggests that cells are selectively dying, and further predicts that NCCs are a developmental dead end. Consistent with the model, NCCs are not detectable after E5, thus excluded from the developmental process (Figures 1A and S1A).

Transcripts of Older transposable elements mark committed cells of ICM

Both of the above models (selective purging and removal of excess cells) predict that the mutagenic young TrEs should not be massively expressed in the cell types that have a developmental future. In committed cells of ICM there are indeed no significant level of young TrEs expressed (Figure 2B). Consistently, we also observe the expression of various APOBECs and IFITM1, implicated in host defence, controlling Young retroelements (e.g. LINE-1) and retroviruses (Grow et al., 2015; Knisbacher et al., 2016)(Figure 2E). In ICM, instead of Young elements, we observe abundant transcripts of various ancient, transpositionally inactive endogenous retroviruses (ERVs), dominantly represented by their full-length versions: LTR2B-ERVL18, LTR41B-ERVE_a, LTR17-ERV17, LTR10-ERVI, MER48-ERVH48, and LTR7-(H)ERVH in ascending order of average expression (Figure 2B). The most robustly expressed is LTR7/HERVH, having a strikingly antagonistic pattern to transcription of young TrEs (e.g. SVA-D and LTR5_Hs) in committed vs non-committed cells (NCCs) (Figure 2F). In EPI, by contrast to PE, the expression of LTR7/HERVH stays high (Figure 2F). Thus, besides marking committed cells and driving a regulatory network of pluripotency (Wang et al., 2014b), LTR7/HERVH might also contribute to lineage determination.

The mutual exclusion of Young and Old TrEs, the former being seen in NCCs, the latter in ICM, could be owing to some third-party switches. Alternatively, activity of one might suppress the activity of the other. As HERVH expression is expected to decline following implantation, knocking down (KD) HERVH in hPSCs can model certain aspects of this developmental stage, when cells discontinue to self-renew and commit to differentiate (Lu et al., 2014; Wang et al., 2014b). The transcriptome of KD-HERVH_h1 cells (Lu et al., 2014) reveals upregulation of Young TrEs (Figure S3E). Thus, we speculate that the future viability of cells with a potential developmental fate is possibly dependent on HERVH involved in suppressing the activity of potentially mutagenic Young TrEs. Hence, activation of Old TrEs (e.g. HERVH) is probably involved in the maintenance of genome stability.

Both ICM and EPI are pluripotent in humans, but only EPI has self-renewal potential

Understanding phylogenetic similarities and differences of the blastocyst would be also important to define a cell type that would be a good candidate for a laboratory model pluripotent cell population. Optimally, in addition to pluripotency, such a cell population would need self-renewal ability, be relatively homogeneous and genetically stable. Both EPI and the newly resolved ICM are potential candidates, as analysis of pluripotency-specific markers (e.g. NANOG, KLF4, POUF5F1/OCT4, etc.) reveals no differential gene expression between EPI and ICM (p-value insignificant) (Figure 3A), arguing against EPI being the only pluripotent cell population in the pre-implantation embryos (Brons et al., 2007).

Nevertheless, the EPI cells can be uniquely characterized by their low cell-to-cell variation, clustering together from both E6 and E7 (Figure 1A). Thus, these cells stably maintain their transcriptome in the blastocyst (a feature of self-renewing cultured cells). To dissect the underlying transcriptional differences between EPI and ICM, we used the top 1217 most variable genes (MVGs). This strategy revealed two distinct clusters of ICM (n=75) and EPI (n=53) on PCA (Figure 3B), and identified 22 and 9 genes, whose exclusive expression distinguishes ICM from EPI, including BMP2, FOXR1, NANOGNB (a duplicated version of NANOG) and NODAL, and LEFTY2 respectively (Figures 3C). Importantly, among the top markers of EPI, NODAL and GDF3 (rank 1 and 2 in our analysis) (Figures 1A and S1E) are implicated in triggering the self-renewal cascade (Niakan and Eggan, 2013), indicating that self-renewal might be a key feature of EPI. Indeed, expression profiling of these markers and further self-renewing genes (e.g. LEFTY1/2, TDGF1, SMAD1) indicates that the self-renewal potential is a property of EPI, but not ICM (Figure 3D). These results define EPI as a pluripotent cell population with self-renewing ability, being a most appropriate candidate for *in vitro* work. In this regard, human EPI is also phylogenetically distinct.

The transcriptomes of ICM/EPI evolve faster compared to the rest of the blastocyst in primates

The above results suggest that the human early embryo is classical in having ICM and step-wise development, but distinct in having a self-renewing EPI. To approach the problem of uniqueness more generally, we ask whether the trancriptomes of the different cells types are evolving at different rates and, if not, what underpins any differences. We observe that ICM and EPI trancriptomes form a single cluster in macaque, but segregate in human (Figure S3A-B), suggesting that the functional divergence is the feature of the human pluripotent cells. To compose comparable data for more detailed cross-species analysis, we calculate the scaled expression of *Homo-Cynomolgus* common 16222 genes and merge the data in a single pool. Applying quality control thresholds, we end up with 11043 genes for further analysis. PCA plotting these merged cross-species data kept the PE and TE lineages together, regardless of their phylogenetic divergence, also supported by 1K unbiased hierarchical clustering

(Figures 3E-F). This suggests that their identity is functionally defined and slow evolving. By contrast, the macaque ICM displays some level of transcriptome-wide similarity to the human EPI (Figures 3E-F and S4A), rather than ICM, perhaps following a similar trend of directionality also observed between marmoset ICM and macaque EPI (Nakamura et al., 2016). Upon loading the first four most significant PCs to segregate the cells on two t-SNE dimensions, we see again the divergent behaviour of ICMs, but the most divergent cell populations between human and macaque are the EPIs (Figures 3E-F).

Can we estimate the speed of the divergence? We detect approximately 300 differentially expressed genes (DEGs) upon comparing entire cross-species blastocyst's single cells in a pairwise manner (Figure S4B). However, when we compare the pluripotent ICM/EPI to the entire blastocyst, we find that the number of DEGs is ~ 4 fold higher (300 vs 1116 DEGs (fold change |2|, adjusted p-value < 0.05)) (Figure 3G), suggesting the pluripotent ICM/EPI was subjected to an accelerated evolution compared to the rest of the blastocyst. Among the DEGs distinguishing pluripotency, we established 12 orthologous genes for *Cynomolgus* and 27 for *Homo* that could be considered as markers (AUC > 85%), including the top markers CYP11A1, STRA6 and ABHD12B, SCGB3A2, respectively (Figure 4A). Among the diverged gene ontology categories between *Cynomolgus* and *Homo* we primarily identify *metabolic*, *immune* and *defense* processes (Figure 3G and S4C).

HERVH-remodelled genes are integrated into the regulatory circuitry of self-renewal in EPI

What underpins the divergence of EPI between macaque and human? To address this we take a systems approach to define networks seen in both or either. Due to the low cell-to-cell variation in EPIs, it is possible to investigate the gene co-expression dynamics by calculating pairwise weight correlation network analysis (WGCNA) (Figures S4D and 4B), and to observe significant pairwise ranked correlations (>80%) on scaled data. Consistent with the predicted self-renewing capacity of EPI, the tightly co-regulated genes in both species include NODAL, GDF3, TDGF1 and PRDM14 (Figures 4B and S4D), associated with the regulation of self-renewal.

Beside conserved gene expression, this approach also allowed us to identify genes between *Homo* and *Cynomolgus* whose expression has been shifted from the ICM to EPI (e.g. MT1G and MT1X) or are unique to human EPI (e.g. ATP12A, ABHD12B, SCGB3A2) (Figure 4B). Notably, while ABHD12B and SCGB3A2 are both annotated in *Cynomolgus*, they are remodelled by HERVH only in human in pairwise comparison, and SCGB3A2 is even human specific (Figures 4C-D). Thus, HERVH-remodelled gene products appear to be incorporated into, and predicted to modulate, the regulatory circuitry of self-renewal in human pluripotency.

Furthermore, in our cross-species analyses, in both ICM and EPI, we detect the upregulation of expressed (log2 TPM > 1) neighbour genes (n=90), located at least 10KB downstream of the HERVH locus in the proximity of expressed HERVH loci in human (Wilcoxon test, p-value < 0.0001) (Figure 4E). In contrast, TE and PE neighbours are downregulated or not affected (Figure 4E).

While many HERVH-enforced transcripts may reflect noise, the function of a number of HERVH-derived transcripts and remodelled genes has been confirmed to be functional and affect pluripotency (Durruthy-Durruthy et al., 2016; Loewer et al., 2010; Wang et al., 2014b; Zhao et al., 2007). Notably, while SCGB3A2 shows equal expression in both ICM and EPI, ABHD12B and other HERVH-derived transcripts (e.g. LINC-ROR, LINC00263, ESRG) are expressed more abundantly in EPI (Figure 4F). Furthermore, while TFPI2 is exclusively expressed in ICM, the expression of its HERVH-remodelled paralogue, TFPI has been shifted to both pluripotent cell types of ICM and EPI (Figure S4E). These examples argue for a possible functional diversification of HERVH-enforced gene regulation/transcripts in modulating pluripotency.

Robust divergence of pluripotency following the split of old and new world monkeys due to HERVH enforced expression

As HERVH-enforced gene regulation/remodelling appears to be involved in the evolution of EPI, we further dissected the emergence of the HERVH-driven regulatory network of pluripotency in primates. Additionally, as HERVH was introduced into the primate genome after the old world new world monkey split, we predict that much of the divergence in regulation will be owing to HERVH.

As models of the pluripotent EPI, we use comparable pluripotent stem cells (PSCs), established from human and various Non Human Primates (NHPs). To determine differentially expressed genomic loci between human and NHPs transcriptomes, we include male PSCs from human, chimp, bonobo (Marchetto et al., 2013) and our own *Gorilla* data (Wunderlich et al., 2014). We additionally generate RNASeq data from comparable *Callithrix* (Muller et al., 2009), where HERVH is not present (Izsvak et al., 2016) as a control. We also extract HERVH-governed genes defined as those differentially regulated in the knockdown cells (HERVH-KD) in the human embryonic stem cell line ESC_h1 compared to a control (GFP-KD) (Lu et al., 2014).

Our cross-species mapping demonstrates how dramatically the expression of human EPI markers (e.g. including LEFTY1/2, NODAL) changes between human and *Callithrix* PSCs (Figure 5A), supporting the dramatic restructuring of the pluripotency network after the split of new world (NWM) and old world monkeys (OWMs). The divergence of PSC transcriptomes is also high among OWMs (Figures S5A-B). Compared to humans, we observe 2340, 375, 172 and 81 differentially expressed genes (DEGs) unique in *Callithrix*, *Gorilla*, *Chimpanzee* and *Bonobo* PSCs, respectively, whereas only 82 genes are shared between them (Figure S5B). The number of unique DEGs is also directly proportional to the total number of DEGs, and the degree of transcriptome diversity agrees with the predicted evolutionary path as inferred from clustering fold change values of all observed DEGs (Figures S5C-D). As we expected, the most contrasting transcriptional pattern is observed between the pluripotent cells of NWM and OWMs.

Next, we determine the expression of human HERVH loci by mapping reads from the comparator species against the human genome, and calculating the level of relative transcription at each locus. Differences between the NWM and OWM are also reflected in gene loss/gain expression events. Remarkably, the major gain (19) and loss (29) events in regulating pluripotency between human and *Callithrix* are due to HERVH-governed gene expression (Figure 5C), underpinning the centrality of HERVH to pluripotency. Among those genes whose expression has tuned down, we identified NR2F2, whose repression was reported to enhance PSC reprogramming in human (Hu et al., 2013) (Figure 5C). Curiously, PRODH is also among the HERVH-controlled gained genes, suggesting that PRODH is under a dual HERV-governed regulation e.g. LTR5/HERVK and LTR7B in brain (Suntsova et al., 2013) (Figure 5C), respectively.

The emergence of the HERVH-based regulatory network predates the human-gorilla common ancestor

When was the co-option of HERVH initiated? To address this, we employ the gene expression profile of the HERVH knockdown as a surrogate of the ancestral – before HERVH – expression profile. Adding up all observed DEGs in any comparison including those identified in HERVH-KD results in around 1100 genes (FDR < 0.05). Hierarchical clustering applying ranked-correlation on their fold-change values reflects the evolution of the primate transcriptome, and pushes human HERVH-KD_ESC_h1 between *Gorilla* and *Callithrix* (Figures 5B and S5E-F), suggesting that the domestication of HERVH predates the human-gorilla common ancestor.

In order to decipher the transcriptional gain and loss of existing HERVH loci between human and NHPs, we scale

the expression of orthologous loci between human-gorilla and human-chimp. Curiously, nearly half of the expressed full-length HERVH loci appear to be human-specific (Figures 5D). Compared to non-human (non-h) PSCs, we observe heavy loss in the number of HERVH expressed loci in hPSCs, whereas, only a few orthologous loci gained expression (Figures 5D and S5G-H). Notably, at orthologous loci, the HERVH-affected neighbour genes are upregulated in hPSCs, but not in non-hPSCs (Figure 5E), suggesting that the robust HERVH-mediated transcriptional control over neighbour genes, thus the modulation of pluripotency occurred quite recently. Upon comparing the orthologous transposable element (TrE) loci between primate species, we notice a marked reduction of overall TrE expression (including the Young, mutagenic elements) in the human pluripotent state (Figure 5F).

Human naïve cultures are heterogeneous and are both evolutionarily and developmentally "confused"

Above we have characterised the various cell types in human pre-implantation embryos, suggested that EPI is the best model to mimic *in vitro* and that much of the circuitry is lineage specific. In particular, the HERVH driven transcriptional network has significantly modulated pluripotency during primate evolution. How well do current *in vitro* pluripotent stem cell cultures match these features? We examine human naïve cell cultures (e.g. 3D morphology) that are either converted from primed cells (Pastor et al., 2016) (e.g. 2D morphology) or freshly established from the human blastocyst (Chan et al., 2013; Pastor et al., 2016; Takashima et al., 2014) and compare them and to their potential primed counterparts.

Upon surveying expression of lineage-specific markers of the blastocyst (AUC cut-off > 85%) we observe that while the naïve cultures upregulate ICM/EPI specific markers and downregulate PE-specific markers when compared to their primed counterparts, they also upregulate NCC markers (Figure 6A). To better profile them, we thus examine genes that are significantly upregulated in at-least 80% of the studied naïve lines when compared to their primed counterparts. We intersect the resultant genes with our lineage specific markers (AUC cut-off > 0.85) (Figure 6B). These *in vitro* cultures appear to represent heterogeneous mixture of cell types in various degree, expressing a diversity of human pre-implantation embryonic lineage markers, including 8-cell, morula, NCC and PE (Figures 6A-B and S6A-B). Thus, although a fraction of cells in naïve cultures resemble real stages of development, the cultures exhibit a non-stereotypical idiosyncratic expression profile and are in this sense "confused". The expression of potentially mutagenic TrEs (expression which may under normal circumstances

lead to apoptosis and a sideways move into the NCC category) might be of concern for the application of these naïve cultures.

Are the transcripts that mark the human-specific features of pluripotent cells *in vivo* well represented in naïve cultures? To address this, we also compared fold-change expression of naïve vs primed cells with the fold-change expression of pairwise human and macaque blastocyst stages (Figures 6C-D and S6C). Our analysis reveals that the human pluripotent blastocyst features are significantly downregulated in naïve cultures (Figure 6C). Similarly, transcripts of those genes that mark the human specific vs primate features of pluripotent stem cells are underrepresented in naïve cultures (Figures 6D).

The naïve cultures are also unusual in having more frequent generation of chimeric transcripts compared to their primed counterparts. These transcripts deriving from two physically independent genomic loci are abundantly generated during early embryogenesis (morula and before), but their generation gradually decays to an approximately steady-state frequency after morula (Figure 6E). Chimeric transcripts are still expressed upon converting naïve cells to primed state (Figure 6F), and might help to explain why these cells are less capable of proper conversion (Pastor et al., 2016).

If EPI is possibly the best cell type to mimic, what genes should the optimal cell type express or not express? To this end, we propose a checklist that could be used to guide *in vitro* studies. We presume that the key properties of the pluripotent EPI, including self-renewal and homogeneity, make this the best candidate to mimic *in vitro*. The checklist hence includes top genes that appear to have a unique expression status in human EPI against the rest of the clustered cells (Figure 6G). We also provide a checklist to exclude genes that do not feature in any real developmental stage in human (Figure 6H). The expression of these genes could induce various aberrant processes that could compromise pluripotency. The checklists include ESRG, LEFTY1, HHLA1 and excludes H19 and KLF2/17 expression (Figures 6G-H). Notably, ESRG and HHLA1 genes carry full-length HERVH sequences, suggesting that human pluripotent cultures should reflect the species-specific features properly, including the expression of HERVH-remodelled genes that have been contributed to fine-tune pluripotency regulation in humans.

Discussion

To discern the uniqueness (or lack thereof) of human early development, here we aimed to decipher the evolution of the blastocyst, with a special focus on pluripotency regulation in primates. In agreement with the hourglass model of development (Kalinka et al., 2010), we find that multiple features of early development are massively divergent between humans and non-human primates. Nevertheless, and despite of the general belief, we propose that the trajectories and the cell types in the primate blastocyst are fundamentally conserved. Our comparative single cell high-resolution analysis of human vs macaque blastocyst can clearly identify the human ICM (marked by NANOG, POU5F1 expression), disproving down the notion that it might not exist in human embryos (Nakamura et al., 2016). Our strategy also helped us to identify phylogenetic differences of pre-implantation embryogenesis in primates.

A distinctive feature of human pre-implantation development is the presence of dynamically changing, hard-to-catch transitory zones between stages. Although ICM is short-lived before it segregates to EPI and PE, it is clearly identifiable, supporting the hypothesis that blastocyst formation, comparable to other mammalian species, occurs in well-defined sequential steps. Upon unmasking ICM, we identified a relatively large (1/3 of E5 cells), previously unrecognised cell population during blastocyst formation. These cells derive from morula, exhibit a high degree of transcriptome heterogeneity (i.e. variation between cells within a resolvable cluster) and fail to express lineage markers. These non-committed cells (NCCs) are subjected to programmed cell death and don't persist after E5.

Apoptosis is thought to play an active role throughout the developmental process. Apoptotic cell are first detectable immediately following embryonic genome activation (EGA), but their timing varies among different mammalian species (Braude et al., 1988). In the mouse, the major activation event occurs during the 2-cell stage (Flach et al., 1982), whereas in humans it occurs later, between the 4- and 8-cell stages (Braude et al., 1988). EGA is accompanied with a global epigenetic change that also de-represses transposable elements (TrEs) (Rowe and Trono, 2011). In human, the expression of mutagenic phylogenetically young elements peaks in morula (Romer et al., 2017; Theunissen et al., 2016). Upregulated Young TrEs are likely to contribute to the observed high heterogeneity of cells segregating from morula, giving rise to both committed (progenitor) and non-committed cells (NCCs). It is parsimonious to suppose that apoptosis is a means to enforce a selective filter against damaged cells that emerge after EGA and fail to properly express lineage markers. NCCs are not observed in mice, perhaps due to the earlier timing of EGA, while they might exist in primates, but were not detected (e.g. selective cell extraction method, (Nakamura et al., 2016)).

According to our quality control model, pre-implantation development is not absolutely directional, but includes a selection process. The boost of TrE activity might determine the fate of the embryo, whether it proceeds to the blastocyst stage or be selected out in a process reminiscent of attrition in fetal oocytes (Malki et al., 2014), involving programmed cell death. Nevertheless, it is likely that only highly damaged cells are filtered out, and some TrE activity is tolerated, resulting even in heritable TrE insertions (van den Hurk et al., 2007). In this scenario, a heterogeneous cell pool of committed cells with a slightly modified genome/transcriptome could be even beneficial. An alternative possibility is that NCCs express mutagenic TrEs because they have no developmental fate. In this model, the upregulation of young TrEs could be a mechanism to lead to the destruction of cells that, for whatever reason, have failed to commit or that are surplus to requirements. Consistent with both models, NCCs are not detectable after E5, thus excluded from the developmental process.

Pluripotency has been evolved in conjunction with host defence

Our study suggests that the evolution of pluripotency in primates primarily affected metabolism, innate immunity and defence response. The connection between the evolution of pluripotency and self-defence has been suggested before (Grow et al., 2015; Wang et al., 2014a) to explain why the human defence response was capable of dealing with the multiple waves of viral invasion during primate evolution, and successfully attenuated Young TrEs (Friedli and Trono, 2015). From the arsenal of complementary processes regulating TrE/viral activities we have detected various APOBECs (Knisbacher et al., 2016) and IFITM1, specifically expressed in ICM vs NCCs. Intriguingly, by contrast to NCCs that are marked by Young TrEs, progenitor cells that passed quality control and continue to participate in the developmental program characteristically express ancient, dominantly full-length ERVs, primarily HERVH. The phenotype of Young TrE activation can be reproduced in HERVH-knockdown conditions in pluripotent stem cells (Lu et al., 2014), suggesting that Old TrEs might be involved in Young TrEs suppression.

The accelerated evolution of ICM/EPI compared to other lineages of the blastocyst refines the hourglass model (Kalinka et al., 2010). The functional divergence between ICM and EPI is a phylogenetically young phenomenon (e.g. not observed in macaques). The pluripotent EPI is the fastest evolving cell types of the blastocyst. While, both ICM and EPI are pluripotent, only EPI forms a self-renewing cell population in humans. Cells of the human EPI are characterised by relatively homogenous transcriptomes, maintained throughout E5-E7, an ideal cell type for *in vitro* culturing. Domestication of HERVH appears to be central to the evolution of pluripotency. ICM and EPI selectively express various HERVH-enforced transcripts, several of them reported to regulate pluripotency

(Izsvak et al., 2016; Loewer et al., 2010; Lu et al., 2014; Ng et al., 2012; Wang et al., 2014b). The features of lineage specification following HERVH invasion in new world monkeys can be further dissected using primate PSC models of the pluripotent EPI. We find that the major gene expression gain (19) and loss (29) events in regulating pluripotency between human and new world monkeys *(Callithrix)* are due to the HERVH-governed gene expression. This involvement of HERVH underpins the great majority of the gene-level differences in genetic architecture between otherwise similar cell types across species. Nonetheless, the major HERVH-driven remodelling of EPI has occurred quite recently, following the split of the *Gorilla*-human common ancestor. Thus, pluripotency of the human form appears to be specific to us humans.

Besides, fine-tuning human pluripotency, HERVH also contributes to the regulatory network of self-renewal (not reported *in vivo*), by incorporating HERVH-remodelled genes (e.g. ABHD12B and SCGB3A2) into the circuitry. Furthermore, HERVH might be involved in cell fate determination (e.g. EPI vs PE). In sum, HERVH appears to selectively marking the genetically stable, self-renewing, pluripotent cell population during blastocyst formation. This profile is consistent with a model in which TrEs are under selection to attempt to control cell fate to promote their own "ends", and thus to manipulate pluripotency and steer cell fate towards germline, within which transposition events have an evolutionary future (Izsvak et al., 2016).

Lessons for *in vitro* models of early embryogenesis

In the last five years, numerous attempts have been reported to derive human naive cells. The quality of these cells has been extensively discussed (Theunissen et al., 2016). Researchers aim at establishing pluripotent stem cell cultures, mimicking the pluripotent blastocyst as closely as possible, and find a suitable non-human primate host for human pluripotent stem cells. *Cynomolgus* is assumed to have both a comparable pluripotency regulation and serve as a potential *in vivo* model of human biology (Dodsworth et al., 2015).

Our analysis holds lessons for establishing self-renewing, pluripotent stem cell cultures *in vitro*. The human pluripotent EPI forms a relatively homogenous cell cluster, has self-renewal capacity and attenuated Young TE activity, potentially making a good choice for in vitro culturing. While, the studied naïve cultures have improved presentation of ICM/EPI markers and underrepresent PE markers, they are heterogeneous, each consisting of a mixture of cells of various identities, and appear to be both evolutionarily and developmentally "confused". Beside EPI-like cells, naïve cultures contain large number of NCCs, as well as cells that display similarity to various embryonic cell lineages, and to pluripotent cells of NHPs instead of human ones. Whether the human-specific

features of EPI could explain difficulties of chimeric studies is yet to be clarified. Furthermore, as the desired end-point of the naïve-like cultures is as a model for early human develop or for transformation into any number of alternative cell types for therapeutic application, it is a matter of substance to discern whether potentially damaging transposition is happening. Filtering out NCCs, preserving human- and lineage-specific features should help to improve derivation and maintenance of human naïve stem cell cultures with improved homogeneity, pluripotency and genome stability. While several strategies have been suggested to establish naive-like cultures, employing a LTR7/HERVH reported-based approach (Wang et al., 2016; Wang et al., 2014b) targets to identify pluripotent EPI-like cells, the only pluripotent, self-renewing cell type in the pre-implantation embryo could have multiple advantages.

Methods

Bulk RNAseq

Data generation:

Total RNA was extracted from *Callithrix jacchus* (Muller et al., 2009) and Gorilla PSCs (Wunderlich et al., 2014) using trizol RNA Mini Prep kit (Zymo research) following the manufacturer's instructions. After extraction, a DNase treatment was applied using TURBO DNA-free Kit (Ambion). The RNAseq library preparation followed the Illumina TruSeq Stranded mRNA Sample Preparation Kit protocol on Illumina HiSeq machine with paired-end 101 cycles.

Data analysis

RNAseq reads with MAP quality score < 30 were removed. We also truncated 2nt from the end of sequencing reads, since their average quality score was not same as the rest of nucleotides. This resulted at least 70 million reads per sample. Next, reads were mapped over the reference genome (Human hg19/GRCh37) and transcriptome model (hg19.refseq.gtf), downloaded from USCS tables (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/). Reads were mapped to their respective reference genomes i.e. human (hg19), chimp (PanTro4), gorilla (gorGor4), marmoset (calJac3) and mouse (mm10) using STAR with our defined settings i.e *–alignIntronMin 20 –alignIntronMax 1000000 –chimSegmentMin 15 –chimJunctionOverhangMin 15 –outFilterMultimapNmax 20* folllowed by constructing STAR genome/transcriptome indices providing their respective RefSeq gtf annotations. As per STAR default we permitted at most two mismatches. We obtained uniquely mapped read counts using featureCounts (Dobin et al., 2013) at gene level with

refSeq annotations. Gene expression levels were calculated at Transcript per million (TPM) from counts over the entire gene (defined as any transcript locating between TSS and TES). This we did using our in-house R script.
The read counts were calculated with *featureCounts* from *subread* package (http://subread.sourceforge.net/), FPKM was calculated using *bamutils* (http://ngsutils.org/modules/bamutils/count/). In order to calculate differential expression at gene level, we used the published model of GFOLD algorithm, which calculates the normalization constant and variance to extract fold changes from unreplicated RNAseq data.

Cross-species analyses

Genes that are differentially expressed (DEGs) between species were obtained by cross-species mapping of RNAseq reads. Reads mappable on both comparators were further mapped on human genome (hg19) using STAR. Cross-species read counts, FPKM and effective fold change was calculated using GFOLD (Fan et al., 2016) on obtained replicated and unreplicated datasets. We mapped human and non-human iPSC RNAseq reads against the human reference genome and gene models to determine the expression level of human genes and repeat elements in NHPs.

Single cell RNAseq data processing

PCA

To define the set of discriminating genes, we calculated the z-score for each gene in each sample in the data frame of all genes across all the samples (i.e. for human 15501 genes in 1285 samples). Each gene was then represented by an across sample vector of z-scores. We then determined the mean of this value across all the genes within the cluster. Those clusters showing a mean(log(Variance/mean)) > 1 were considered as most variable clusters. All the genes in these clusters were considered as most variable genes (MVG). The above PCA analysis clearly resolves the merged datasets of Oocytes to blastocysts with embryonic days E3 and E4. However, E5 onwards appears as an unresolved cloud. In order to resolve this cloud, we first ran t-SNE on single cell data for E5 stage. As this resolved the stages we were interested in, we then applied t-SNE to the full dataset, enabling full resolution of discrete stages in early human development.

t-SNE

We used Seurat (http://satijalab.org/seurat/) and SCDE (http://hms-dbmi.github.io/scde/) packages from R. A R package 'Seurat' was used to obtain most variable genes, markers for given clusters, most significant principle components, t-SNE analysis and visualizations. Samples expressing more than 5000 genes and genes that express

(Log2 TPM > 2) in at least in 1% of total samples were subsequently selected for analysis. This resulted in 1285 single cells carrying expression levels of 15501 genes for human E3-E7 samples. We separated cells by applying most variable genes ({log(Variance) & log(Average Expression)} > 2) to the dimension reduction methods, notably principal component analysis (PCA) and t-stochastic neighbor embedding (t-SNE). Briefly, We reduced the dimensionality of our dataset using principal component analysis. As previously described in (Macosko et al., 2015), we ran PCA using the '*prcomp'* function in R, and then utilized a modified randomization approach ('*jack straw*'), a built-in function in "Seurat" package to identify 'statistically significant' principal components in the dataset. This approach gave us 9 significant principle components (PCs) for E3-E7 stages, 5 significant PCs for E3-E4 stages and 6 significant PCs for E5 stage. Using these cell loadings for significant PCs of respective analysis, we applied *t*-distributed stochastic neighbor embedding (*t*-SNE), a machine learning algorithm for clustering the single cells to visualize the data in two dimensions. This approach illustrated 10 clusters from E3-E7, 3 clusters each for E3-E4 and E5 cell populations. A gene qualified as a marker of a given cluster if it fulfilled three criteria: the gene must be overexpressed in that particular cluster (average fold difference > 2 compared to the rest of clusters), must also be expressed (Log2TPM > 2) in at least 70% of cells in that particular cluster and Area Under Curve (AUC) value must be greater than 80%.

Analysis of repetitive elements

To estimate the expression level for repetitive elements on their locus, we used two strategies. The long reads in (Yan et al., 2013) data allowed us to cover and calculate CPM or RPKM for unique loci of TrEs. In contrast, data from (Petropoulos et al., 2016) was suitable only to detect the average expression of any given TrE family. For this analysis, we considered multimapping reads only if they were mapping exclusively within a TrE family. Than counted one alignment per read to calculate counts per million (counts normalized per million of total reads mappable on human genome). Note that datasets from different layouts (single vs bulk RNAseq) were never merged into one data frame to perform TrEs comparative analysis, as no any normalization method was effective enough.

Homo-Cynomolgus

For this analysis, we selected cells from the pre-implantation blastocysts of human (Petropoulos et al., 2016) with 228 cells (ICM, EPI,PE and TE) and Cynomolgus (Nakamura et al., 2016) with 170 cells (ICM, EPI, Hypoblast and TE). For cross-platform single-cell RNAseq data, counts were merged on gene names, log2 TPM+1 was calculated in similar way as mentioned above. We redefined ICM, EPI, PE and TE cells using only those genes that were

annotated in Refseq gene track of both human and *Cynomolgus* species. We checked the validation of this analysis by visualizing the selected gene expression (log TPM values) of conserved lineage markers across vertebrate blastocysts (Nakamura et al., 2016). Plots shows a similar expression pattern of e.g NANOG, POU5F1, ICM/EPI; SPIC, ICM; NODAL, GDF3 and PRDM14, EPI; APOA1, GATA4 and COL4A1, PE; DLX3, STS and PGF marking TE in both human and macaque. We then, filtered out those genes that are not annotated in any of the given species. This resulted in 16222 individual genes that were merged in single pool. In total, 11053 orthologous genes are analysed that are expressed in any of 5 cells. Variation due to batch effects was adjusted using COMBAT (Johnson et al., 2007) from R package sva. We checked the normalization status by drawing PC biplots using various subsets of clustered genes. This assured us that cells did not cluster on the basis of platform or species.

Self-renewal regulatory network

We created a data frame of single-cell data for ICM, EPI, PE and TE from days E5, E6 and E7, carrying expression values (TPM) of all Human MVGs. We then computed pairwise Pearsons correlation for all MVG. We then selected only those paired genes that show strong correlation or anti-correlation (threshold rho > |0.80|), as shown in heatmap (Figure S4f). A network was constructed on genes showing the highest level of ranked correlation among each other, with rho > 0.80, using igraph (http://igraph.org/r/) package from R. Arrows show the linking (links based on a preset level of preferential attachment (Barabasi-Albert model)) between genes. The direction of the arrows is manually set under the criterion that from genes appear first in human pre-Epi, to genes appearing next. The size of a circle represents the number of instances a gene is upstream (nodes) of its paired partners (edges). Genes in the network are markers of human EPI and colors are assigned as to their expression, or lack thereof, in mouse embryogenesis.

Visualization of reads

Mapped reads from single cell transcriptomes of human embryonic development were merged for each stage, defined as EPI, PE and TE (Yan et al., 2013) using the markers shown on Figure S1E. Mapped reads in bam format were converted into bedGraph format to visualize through IGV over Refseq genes (hg19). Conservation track was visualized through UCSC genome browser under net/chain alignment of given non-human primates (NHPs) shown in Figure 4D and, later on, merged beneath IGV tracks.

Pathway analysis of differentially expressed genes

Canonical pathways and biological function of the identified differentially expressed genes in data sets were investigated using KEGG Pathway or Gorilla tool. Overrepresentation of a biological pathway was assessed by Fisher's exact test and corrected for multiple testing by the Benjamini-Hochberg procedure. The ratio (overlap) is calculated as a number of genes from the dataset that map to the pathway divided by the number of total genes included into the pathway.

Detection of chimeric transcripts from RNAseq data

In order to determine chimeric transcripts, we first aligned the reads using universal aligner STAR using the parameters written above that can discover canonical and non-canonical splice and chimeric (fusion) sites. We kept only the junctions that were identified with a minimum of 6 uniquely mapped reads. Any novel genes with resemblance to mitochondrial genes were excluded from the analysis. Either donor site or acceptor site mapping to the mitochondrial genome was considered grounds for exclusion. To exclude chimeras derived from repeated elements, we identified those novel transcripts that had at least 6 consecutive bps from known repeated elements (repeat specified in hg19 rmsk.gtf).

Human embryo manipulation and microscopy analyses

Prior to the start of the project, the whole procedure was approved by local regulatory authorities and the Spanish National Embryo steering committee. Cryopreserved human embryos of the maximum quality were donated with informed consent by couples that had already have undergone an in vitro fertilization (IVF) cycle. All extractions/manipulations were carried out in a GMP certified facility by certified embryologist in Banco Andaluz Celulas Madre, Granada, Spain. Confocal analyses of LINE-1 ORF1p expression were analyzed on a Zeiss LSM 710 confocal microscope using a previously described method (Macia et al., 2017). – Antibodies for the immunostaining: Rabbit anti LINE-1 ORF1p, 1:500, a generous gift of Dr Oliver Weichenrieder (Max Planck, Germany). Secondary antibody: Alexa 488 Donkey anti Rabbit, 1:1000 (Thermo). Mouse anti H2AXγ, 1:200, clone 3F2 (Novus). Secondary antibody: Alexa 555 Donkey anti Mouse, 1:1000 (Thermo). DAPI (Thermo) was used at 1:500.

Acknowledgments

Z.I. is funded by European Research Council, ERC Advanced [ERC-2011-ADG 294742]. L.D.H. is funded by European Research Council, ERC Advanced [ERC-2014-ADG 669207]. J.L.G.P´s lab is supported by CICE-FEDER-P12-CTS-2256, Plan Nacional de I+D+I 2008-2011 and 2013-2016 (FIS-FEDER-PI14/02152), PCIN-2014-115-ERA-NET NEURON II, the European Research Council (ERC-Consolidator ERC-STG-2012-233764), by an International Early Career Scientist grant from the Howard Hughes Medical Institute (IECS-55007420), by The Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund (ISFF2) and by a private donation by Ms Francisca Serrano (*Trading y Bolsa para Torpes*, Granada, Spain).

Author contributions

The authors declare that they have no conflicts of interest. Z.I., M.S and L.D.H. designed the study and drafted the manuscript. M.S. conceived the idea, designed and performed all the computational analyses. S. W. and U. M. have assisted in providing the NHP lines. G.G.S. provided the material isolated from the Callithrix ESC line CJES-001. The official provider of the *Callithrix* Embryonic stem cell line is the Central Institute for Experimental Animals (CIEA), 1430 Nogawa, Miyamae, Kawasaki 216-0001. T.J.W, J.L.C. and J.L.G.-P. carried out all the work with cryopreserved human embryos.

Competing interests statement

The authors declare that they have no competing financial interests.

Supplementary Information is linked to the online version of the paper.

Figure legends

Figure 1. High-resolution dissection of human pre-implantation development

A. Two-dimensional t-SNE analysis of human single-cell pre-implantation transcriptomes using 1651 most variable genes (MVGs) resolves the following distinct cell populations: 8-cell at E3, morula at E4, non-committed cells (NCCs) and transitory cells at E5, pluripotent epiblast (EPI) at E6-E7, primitive endoderm (PE) at E6-E7, mural and polar trophoectoderm (TE) at E7. At E5, cells presenting none of the known lineage markers referred as non-committed cells (NCCs), whereas cells express multiple markers are annotated as transitory cells. The most discriminatory genes of each clusters are listed in boxes. Numbers in brackets refer to AUC values. Colors indicate unbiased classification via graph based clustering, where each dot represents a single cell.

B. *t*-SNE clustering of E5 cells using the first six PCs (Figure S2C) as input loadings reveals three distinct clusters. The analysis dissects the previously unattended cluster (shown in dark red) of non-committed cells (NCCs). Each dot represents an individual cell.

C. Multiple feature plots illustrate the three clusters (shown on Figure 1B), distinguished by their strong expression of known markers ICM (IL6R), EPI (NANOG), PE (BMP2), Pre-TE (DLX3 and TMPRSS2). NCCs are flagged by BIK (BCL2-Interacting Killer/Apoptosis-Inducing NBK) as illustrated in the feature plot. Each dot represents an individual cell. Colour intensity indicates the expression of the marker gene.

D. Heatmap visualization of scaled expression [log TPM (transcripts per million)] values of discriminative set of genes for each cluster defined in Figure 1B (AUC cut-off >0.90). Color scheme is based on z-score distribution from -2.5 (light blue) to 2.5 (purple). Left margin color bars highlight gene sets specific to the respective clusters in Figure 1B and top margin colour bars define the same for cells. ICM specific genes are marked by (*) or (#) are also expressed at E6-E7 in EPI or PE, respectively.

E. Heatmap of the row-wise scaled expression (log TPM) levels of selected marker genes (Figure 1C-D) for Pre-TE (DLX3), EPI (NANOG), PE (BMP2), ICM (IL6R) and NCC (BIK). Colour bars under the dendrogram were set manually showing the clusters of distinct cells expressing differential combination of markers. Transitory cells (T and T1) are co-expressing multiple lineage markers.

Figure 2. Phylogenetically young and old transposable elements are antagonistically expressed in non-committed cells and the inner cell mass

A. Dendrogram based hierarchical clustering using ranked correlation and complete linkage method on averaged expression from the distinct cell populations of human pre-implantation embryogenesis (via bootstrapping (1000 replicates). The transcriptomes of all the distinct populations are pooled together and averaged for the analysis. Height of dendrogram represents the Euclidian distance of dissimilarity matrix, numbers in red and blue indicate au and bp values from the bootstrapping.

B. Scatterplot shows the comparison of normalized mean expression in CPM (Counts Per Million) of various TrE families between averaged pool of ICM (x-axis) and NCC (y-axis) cells. Read counts are normalized to total mapable reads per TrE family. Note: Uniquely mapped reads were considered as one alignment per read. Multimapping reads were considered as one alignment only if they were mapped to multiple loci, but exclusively within a TrE family. Every dot corresponds to a TrE family. TrE families enriched in ICM (red) vs NCC (blue).

C. Representative images show immunofluorescence staining of human early (E5) blastocysts POU5F1/OCT4, nuclear, green; LINE-1 (L1_Hs) ORF1p, cytoplasmic, red; DAPI, nuclear, blue). Note: POU5F1, used to stain ICM is significantly enriched in ICM. See also violin plot (upper left panel) that visualizes the density and expressional dynamics of the POU5F1 compared with TE and NCC at E5. Co-staining demonstrates the exclusive expression of POU5F1 and L1_Hs_ORF1p during the formation of blastocyst (arrows). Note: The cells expressing higher POU5F1 are compacting to form the ICM at polar region of the blastocyst are not stained for L1_Hs_ORF1p. L1_Hs_ORF1p stains scattered cells, not included in the compacted population of cells. (TE cells were not considered for this analysis). L1_Hs belongs to a group of mutagenic, Young TrEs and supports transposition of both L1 and the non-autonomous Alu and SVA elements.

D. Representative images show immunofluorescence staining of two (upper and lower panels) human early (E5) blastocysts. H2AXγ staining (green) visualizes double stranded DNA damage (DAPI, blue). H2AXγ is readily detected in a fraction of cells during blastocyst the formation. Damaged/dying cells accumulate H2AXγ signals. H2AXγ stained cells might still have integrity and exhibit normal oval shape (upper panel) or loose integrity (lower panel). Note: As cells might die upon thawing, we only analysed blastocysts that fully developed *in vitro* from E2 embryos.

E. Violin plots visualize the density and expressional dynamics of various proteins, implicated in host-defence against retroelements and viruses (APOBEC3C/D/G and IFITM1) in non-committed (NCC) vs committed cell populations (pre-TE and ICM) of the E5 human blastocyst. Note: the transcription of the depicted genes mark ICM at E5.

F. Violin plots visualize the density and expressional dynamics of transposable elements, SVA, LTR5_HS (Young) vs LTR7/HERVH (Old) the in non-committed (NCC) vs committed cell populations (PE and EPI) of the human blastocyst. Young elements are enriched in NCCs vs committed cells (p-value < 0.000072), whereas LTR7 and HERVH-int are enriched in EPI vs PE and NCCs (p-value < 0.00037).

Figure 3. Accelerated divergence of ICM and pluripotent epiblast during primate evolution

A. Violin plots illustrate expression distribution of selected genes associated with pluripotency, (p-value > 0.23) genes (grey for EPI; green for ICM).

B. PCA biplot showing the analysis of human ICM (n = 75) and EPI (n = 52) cells using the most variable genes (n=687). PC1 versus PC2 demonstrates the splitting process of ICM to EPI based on transcriptional proximity

between the mentioned lineages; each dot represents an individual cell; coloured legend for each subset is shown on the top of the plot.

C. Heatmap showing scaled expression (log TPM values) of discriminative gene sets defining EPI and ICM (AUC cutoff ≥ 0.85). Colour scheme is based on *z*-score distribution, from –2.5 (blue) to 2.5 (purple).

D. Violin plots illustrate expression distribution of candidate genes associated with self-renewal (p-value < 0.00005) genes (grey for EPI; green for ICM).

E. t-SNE visualization of significant genes contributing to PC1 and PC2 from the cross-species normalized scaled genes (commonly annotated in *Homo* and *Cynomolgus* (Cyno) Refseq gene track format aka gtf) expression (Log2 TPM) estimates in *Homo* and *Cynomolgus* blastocyst single cells aka. ICM, EPI, PE and TE. For input loading the 1055 most variable genes across the merged cross-species datasets were selected. Every dot represents a single cell and colour code for respective cells are pinned next to dots with same colour. While conserved cell population across the species are circled, arrows point to the diverged cell clusters between *Homo* and *Cynomolgus* (e.g. ICM, but mostly EPI),

F. Dendrogram via bootstrapping-based (1000 replicates) hierarchical clustering using ranked correlation and complete linkage method on averaged expression from the cell populations (mentioned in Figure 3F) transcriptome pooled together. Height of dendrograms represent the Euclidian distance of dissimilarity matrix, numbers in red and blue indicate au and bp values from bootstrapping.

G. Volcano plot illustrates the differentially regulated genes (DEGs) between *Homo* and *Cynomolgus* pluripotent states (ICM and EPI). In total, 7583 genes that are expressed in 50% of the cells of any of the two species are plotted (y-axis, log2 fold change calculated from 'seurat' package and adjusted p-value; x-axis, two-tailed t-test which is further adjusted by multiple corrections). 1116 differentially regulated genes are shown as purple dots (fold change |2|, adjusted p-value < 0.05). Enriched gene ontologies of DEGs are indicated in the plot with corresponding p-values.

Figure 4. HERVH-remodelled genes are incorporated into the regulation of self-renewal in primates

A. Heatmap showing scaled expression of discriminative gene sets defining either EPI or ICM (AUC cutoff ≥ 0.85) between *Homo* and *Cynomolgus*. Colour scheme is based on row-wise z-score distribution ranging from –2.5 (blue) to 2.5 (purple).

B. Differential transcriptional network regulating self-renewal between *Cynomolgus* and *Homo* by analysing single cell transcriptomes. Only pairs having a strong correlation with each other are considered (Spearman's coefficient > 80%). Note that the human EPI was the only cell population where, due to low cell-to-cell variation, paired gene expression genes could be obtained. Only genes annotated in both species are considered. For each pair, nodes and edges are decided on their expressional dynamics in the EPI cluster. Size of the nodes is proportional to the number of components the gene expression is paired with in the network. Colours denote species specificity: genes whose expression is shifted from monkey ICM to human EPI (grey); expressed in both human and monkey EPI (blue); in human EPI only (red).

C. Violin plots visualize the density and distribution of gene expression (log TPM values) of selected orthologous genes in the *Homo* vs *Cynomolgus* blasctocyst lineages (left vs right panels). Note: Selected, HERVH-remodelled genes (e.g. SCGB3A2, ABHD12B and HHLA1) are specifically marking the human pluripotent lineages. Expression of ATP12A is shifted from *Cynomolgus* TE to human EPI (not HERVH-dependent).

D. HERVH-enforced gene expression marks distinct stages of early development. Integrative Genome Visualization (IGV) of uniquely mapped reads over a specific gene and the proximal full-length HERVH locus. Arrows show the annotated (black) and HERVH-enforced (purple) transcriptional start sites (TSSs). Transcription skipped at annotated (empty box) and HERVH-enforced TSSs (shaded box) are shown. Both genes loose their annotated TSS and proximal exons to form HERVH chimera. The chimeric ABHD12B transcript is expressed from zygote to EPI, but expression pauses in 8-cell/morula, exons of ABHD12B upstream of HERVH/LTR7 are skipped. While ABHD12B HERVH appears to be intact in Chimpanzee, it has several deletions compared to the human version (not shown). SCGB3A2, implicated in pluripotency, exhibit partially overlapping expression patterns, usage of distinct human-specific HERVH TSS and loss of annotated TSS and proximate exons. Lowest panels show phylogenetic conservation status, the presence (thick line) and the absence (narrow line) of the human sequence compared to the Chimpanzee, Orangutan, Rhesus and Marmoset assemblies.

E. Notched boxplot represents the distribution of average difference (at log2 scale) of LTR7-HERVH neighbour (upto 10 KB downstream, n=53) genes expression. Note: Cell populations are pooled together, scaled and averaged. Only genes, commonly annotated in both human and macaque Refseq gtf were taken for this action. The upregulation of HERVH neighbour gene expression was observed only in ICM and EPI, but not in PE and TE transcriptomes.

F. Violin plots illustrate expression distribution of selected genes remodelled by HERVH (grey for EPI; green for ICM).

Figure 5. Loss and gain of expression in pluripotent cells during primate evolution is mostly due to HERVH

A. Scatterplot shows differential expression of human pluripotent EPI genes (n=308, AUC cut off > 80%) in human and *Callithrix* pluripotent stem cells. Expression values are obtained as RPKM calculated on the human genome from reads mappaple on both genomes. Dots represent genes that have either lost (blue) or gained (dark-red) expression in human PSCs, respectively. Note: We also considered those genes that contained zero mappable reads in either of the analysed species (e.g. ESRG).
B. Barplots combined with dendrograms display the comparison of genes in non-human primate and human pluripotent stem cells (PSCs) controlled by HERVH transcription. Barplots show the number of significant DEGs (FDR<0.01 and fold change > 2 or < -2) of gene lists obtained from *Callithrix* (n=1) vs human (n=4), *Gorilla* (n=2) vs human (n=4), *Chimpanzee* (n=4) vs human (n=4), *Bonobo* (n=4) vs human (n=4), HERVH-KD vs GFP-KD (n=2) in ESC_h1. In case of two replicates, we selected only those genes, which were differentially regulated in both replicates in a similar fashion. Sorted according to HERVH-KD. Dendrogram is calculated by ranked correlation and Euclidian distance method on the expression matrix of most variable genes (n=~2800).
C. Combined barplots show the gain (n=19) and loss (n=29) of gene expression between human and *Callithrix* (no HERVH is present) pluripotent stem cells due to HERVH regulation. Left and right panels show and fold-change |2| values in KD-HERVH vs KD-GFP in ESC_h1s and cross-species expression (RPKM), respectively. Green bars in the right panel are genes expressed significantly in human, but not in *Callithrix* (reads are mapped on the human genome). Left panel shows the same set of genes downregulated when HERVH is depleted using by RNAi (HERVH-KD vs GFP-KD in ESC_h1s). The opposite scenario is shown in purple. (FPKM < 1 was considered as loss).
D. Heatmap displays the dynamic expression of HERVH and HERVK (log FPKM) in primate pluripotent stem cells. Z-scores are calculated on the expression obtained by mapping reads on the human genome from *Bonobo*, *Chimpanzee* and *Gorilla*. The numbers represent expressed genomic loci. Note the lower number of active genomic copies of HERVH in human compared to the rest of the primates.
E. Notched boxplots represent the distribution of LTR7-HERVH affected gene expression (upto 10 KB downstream, n=53) from cross-species mappable reads in various primate pluripotent stem cells (at log2 scale). Hash (#) bars show the pairwise calculation of p-values between human and non-human primates.

F. Boxplots show the pairwise distribution of global genomic expression of transposable elements (TrEs) between human (hg19 version) and non-human primates. We consider only those TrE loci that are mappable in both comparators and are expressed (Log2 FPKM > 1). P-values were calculated by Wilcoxon test.

Figure 6

Current naïve cell cultures don't reflect human uniqueness properly

A. Boxplot shows the upregulation of human pre-implantation lineage markers in various naïve cell cultures compared to their respective primed counterparts (GFOLD calculated on Reset cells/H9_ESCs (Takashima et al., 2014; Theunissen and Jaenisch, 2014), 5iL_SSEA_Neg/UCLA1_primed, UCLA_20n/UCLA_20n_primed, 5iL_SSEA_Pos/UCLA1_primed (Pastor et al., 2016) and Chan_3iL/h1_ESCs (Chan et al., 2013). The marker genes selected for the analysis flag distinct lineages (e.g. NCC, ICM, EPI, PE; n=number of marker genes). The AUC cutoff values were chosen using the following criteria: (i) should be putative markers of any distinct lineage in human pre-implantation embryos (AUC cutoff > 0.85) after EGA; (ii) should be expressed in either naïve or primed cells.
B. Stacked bar plots showing the expression of human pre-implantation lineage marker genes in various naïve and their respective primed cell cultures. Colours and the gene list are as in Figure 6A.
C. Notched boxplot shows the distribution of differential gene expression (log2 fold change) in various human naïve cell cultures with their respective primed counterparts (steel blue boxes). The genes chosen for this analysis (n=246) are commonly upregulated (log fold change > 1) in all lineages (ICM, EPI, PE and TE) of the human blastocysts compared with their counter lineages in *Cynomolgus* blastocysts (gold boxes).
D. Notched boxplot shows the distribution of differential gene expression (log2 fold change) in human naïve cell cultures with their respective primed counterparts (steel blue boxes). The genes chosen for this analysis (n=197) are commonly upregulated (log fold change > 1) in human pluripotent stem cells (hPSCs) vs all analysed non-hPSCs (*Callithrix*, *Gorilla*, *Chimpanzee* and *Bonobo*) (violet boxes).
E. Bar plot showing the number of chimeric transcripts detected in single cell transcriptomes of various human pre-implantation stages (Yan et al., 2013). Note that the generation of chimeric transcripts is tuned down after morula.
F. Bar plot showing the number of chimeric transcripts detected in various naïve cultures, in their respective primed controls and in their converted counterparts. (n=number of samples analysed).

G. Barplot of EPI-like checklist. Log2-fold change values of genes either specifically expressed in pre-EPI/EPI or upregulated in EPI vs rest of the cells (ROC) in the human embryos, and show opposite pattern when naïve cells are compared with their primed counterparts.

H. Barplot of EPI-like checklist. Log2-fold change values of genes either specifically repressed in pre-EPI/EPI or downregulated in EPI vs rest of the cells (ROC) in the human embryos, and show opposite pattern when naïve cells are compared with their primed counterparts.

Supplementary Figures

Figure S1, related to Figure 1. Tracing human pre-implantation embryogenesis

a. Re-ordered phylogenetic tree of clusters shown on Figure 1A. The tree is constructed using *"BuildClustertree"* built-in function of the R package *'Seurat'*. Nodes are shown in grey boxes and numbers are in the order of their position on the tree. Colour codes are as in Figure 1A. Numbers in coloured boxes denote the number of cells in each representative cluster. Note: From the 1,285 single cells in total (Petropoulos et al., 2016), the independently clustering 16 single cells were added to the transitory category from E6 and E7 (originally 410 cells). At E5, a cluster of 99 cells does not express any of the lineage markers (non-committed cells, NCCs) (dark red).

b. Two-dimensional t-SNE analysis of human single-cell pre-implantation transcriptomes using 872 most variable genes (MVGs) resolves 8-cell and morula stages.

c. Defining markers of 8-cell and morula stages. The most discriminatory genes of 8 cell stage and the two distinctive cell populations of morula, marked by LEUTX, GATA3 and HKDC1. Colours indicate unbiased classification via graph based clustering, where each dot represents a single cell.

d. Tracing the human embryonic development progression from zygote to blastocyst. Principal component analysis (PCA) of cross-platform 1285+104 single-cell pre-implantation transcriptome (Petropoulos et al., 2016; Yan et al., 2013) using 1,583 most variable genes (MVGs). Developmental stages defined as in (Blakeley et al., 2015; Petropoulos et al., 2016).

e. Heatmap displaying the scaled expression (log TPM values) of discriminative gene sets (AUC cutoff ≥ 0.90) defining cell populations of morula (n=171), EPI (n=52), PE (n=45) and TE (n=99) reported in Figures 1A and S1A. Heatmap colour scheme is based on z- score distribution from -2 (light blue) to 2 (purple). Note: In the previous study (Petropoulos et al., 2016), EPI, PE and TE genes were defined as differentially expressed at defined embryonic days (E5, E6 and E7).

Figure S2, related to Figure 1. Dissection of the human blastocyst formation

- **a.** PCA of E5 cells (n=300) using most discriminating genes (n=526) reveals three clusters on PC1 and PC2.
- **b.** Scatterplot displays the comparison of averaged gene expression of NCC and ICM cells pooled together in pairwise manner. Red and blue dots represent genes whose expression is enriched in either ICM or NCCs. Annotated genes are: unchanged housekeeping genes (GAPDH and ACTB), genes enriched in either NCCs or in ICM.
- **c.** Heatmaps showing scaled expression (log TPM values) of top 30 discriminative genes per PC (Principle Component) in E5 cells visualized on the first six most significant components. These six PCs were used as input loading for further t-SNE analysis of E5 cells. This analysis explores the pairwised correlated and anti-correlated gene sets among E5 cells.
- **d.** Venn diagram shows the top 3 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enriched either in ICM (red) or in NCC (blue) and shared between them (merged, grey). Numbers in bracket indicate the number of genes involved in the respective pathways. Black brackets show ICM numbers slashed by NCC ones.
- **e.** Multiple violin plots visualize the density and distribution of gene expression (log TPM values) of selected genes that are upregulated in human NCC vs EPI/PE. The depicted genes are involved in regulating apoptotic pathways (KEGG: hsa04210, Gene Ontology GO:008219, GO:0012501 and GO:0006915) (Wilcox test, p-value < 7.135e-06).
- **f.** Schematic of stepwise progression of blastocyst formation based on tracing the marker from cell-to-cell at E5. Bracketed numbers indicate the number of cells showing the characteristics of the various cells types.

Figure S3, related to Figure 2. Non committed cells upregulate apoptotic gene expression and Young transposable elements

a. PCA of the distinct lineages of the human blastocyst by most variable genes among the shown groups (228 cells, 1055 genes). a-b: Note: We considered ICM, EPI, PE and TE cells and only those genes were taken into account that were annotated in Refseq gene track of both human and *Cynomolgus* species. Every dot represents single cell. Colours flag distinct cell types in the human blastocyst.

b. PCA of the distinct lineages of the *Cynomolgus* blastocysts by most variable genes among the shown groups (170 cells, 1237 genes).

c. Multiple violin plots visualize density and distribution of selected gene expression (log TPM values) of conserved lineage markers across vertebrate blastocysts (Nakamura et al., 2016). Plots shows a similar expression pattern of e.g NANOG, POU5F1, ICM/EPI; SPIC, ICM; NODAL, GDF3 and PRDM14, EPI; APOA1, GATA4 and COL4A1, PE; DLX3, STS and PGF marking TE in both human and macaque.
d. Expression dynamics of L1-ORF1 (red) by immunofluorescence staining during the formation of the blastocyst. Note that L1_HS_ORF1p accumulates in the cytoplasm of Pre-TE and in the blastocoel cells (DAPI, blue). See also Movie 1.
e. MA plot displaying the comparison of average difference of normalized expression (CPM) of various transposable element (TrE) families in knockdown HERVH vs control cells (Lu et al., 2014). y-axis, Logfold change GFP_KD (n=2) vs HERVH_KD (n=2) in ESCs_h1; x-axis, average expression of TrE families in the same dataset. Dots represent TrE families downregulated (blue) or upregulated (red) (log2 Average CPM > 5 and average difference > 2) in HERVH_KD ESC_h1.

Figure S4, related to Figure 4. Divergence of the pluripotent cell types in human

a. PCA visualization of significant genes contributing to PC1 and PC2 from the cross-species normalized scaled genes. Only genes commonly annotated in both human and *Cynomolgus* Refseq gene track format aka gtf) are considered. Expression (Log2 TPM) estimates are shown in *Homo* and *Cynomolgus* blastocyst single cells aka. ICM, EPI, PE and TE. The 1,055 most variable genes (as in Figure 3F) across the merged cross-species datasets were selected for input loading. Every dot represents a single cell.
b. Histogram showing the distribution of differentially expressed genes (DEGs) between human and macaque blastocyst. In total, 11,053 orthologous genes are analysed that are expressed in any 5 cells. The analysis detected 181 down-regulated, 141 upregulated genes in the human blastocysts (p-value < 0.05 and log2 fold change > 1). The differential expression of further 2226 genes was not significant.
c. Table displays the gene ontology of DEGs between human and macaque blastocyst. Analysis was performed using Gorilla tool.
d. Heatmap of a correlation matrix visualizing the pairwise correlation of most variable gene expression in single cells of human blastocysts lineages. For the analysis 1076 genes were used whose dynamic expression was suitable to segregate distinct blastocyst lineages on first two principal components (Figure S3A). K-means clustering provided three major clusters of genes marking TE, ICM and EPI (from left to right). Framed box contains a list of tightly correlated EPI markers. Networks beneath the correlogram illustrates the paired pluripotent genes with ABHD12B and SCGB3A2 calculated with

similar parameter shown in Figure 4C.

e. Violin plots visualize the density and distribution of TFPI and TFPI2 (the paralogue of TFPI) expression (log TPM value) in EPI (grey) and ICM (green). Note the differential/exclusive gene expression of the TFPI paralogues in the pluripotent ICM and EPI cells. Each dot represents an individual cell.

Figure S5, related to Figure 5. The robust divergence of pluripotency regulation in primates is HERVH-enforced

a. Boxplots show the normalized global expression estimates of RNAseq datasets from various primate induced pluripotent stem cells (PSCs) analysed in this study (cross-species mapped).
b. Venn diagram displays the divergence of PSC transcriptomes in primates (e.g. human, *Bonobo*, *Chimpanzee*, *Gorilla* and *Callithrix*). The numbers in the Venn diagram denote differentially expressed genes (DEGs) (FDR<0.05 and fold-change |2|). RNAseq data of various non-human primates were compared to human PSCs. Only cross-species reads mappable to both genomes were considered. Gene expression was calculated on the human genome, using the human gene model.
c. Barplot showing number of DEGs calculated from reads mappable to both comparators.
d. Heatmap showing level of differential expression of cross-mapped genes between human and non-human primate PSCs. The analysis included all the differentially expressed genes in any of the given comparisons (as in Figure S5C).
e. The impact of HERVH-mediated regulation on the evolution of primate pluripotency. As in Figure S5B, but also including DEGs genes upon HERVH knockdown (HERVH-KD in ESC_h1 vs GFP-KD in ESC_h1).
f. Barplot showing number of DEGs affected by HERVH expression. As in Figure S5C, but also including DEGs genes upon HERVH knockdown (HERVH-KD in ESC_h1 vs GFP-KD in ESC_h1).
g. Heatmap displays the loss and gain of expression of orthologous HERVH loci between human and gorilla PSCs. Only RNAseq reads, mappable to both human and *Gorilla* reference genomes were considered.
h. Heatmap displays the loss and gain of expression of orthologous HERVH loci between bonobo, chimpanzee and human PSCs. Only RNAseq reads, mappable to both human and *Chimpanzee* reference genomes were considered.

Figure S6, related to Figure 6. How faithful a primate's developmental model could be for human pluripotency?

Multiple heatmaps of pairwise Spearman's correlation showing purple as positive and light blue as negative correlation of four subsets of genes obtained from k-means clustering the transcriptome. Dendrogram shows the clustering of samples based on Euclidian distance of dissimilarity matrix using 1210, 1265, 1181 and 2214 genes from cluster 1 to cluster 4. Note: Based on the input gene sets *in vitro* naïve cultures cluster with NCC and Morula or ICM and EPI. Reset /H9ESCs (Takashima et al., 2014), 5iL_SSEA_Neg/UCLA1_primed, UCLA_20n/UCLA_20n_primed, 5iL_SSEA_Pos/UCLA1_primed (Pastor et al., 2016) and Chan_3iL/ESC_H1s (Chan et al., 2013).

a. Boxplot showing the upregulation of various early embryonic lineage markers (shown in the table next to the boxplot with respective AUC cutoff values) in cultured naive cells compared to their respective primed counterparts (GFOLD calculated on Reset cells/H9ESCs). The selected genes were chosen according to the following criteria: (i) should be putative markers of any distinct lineage in human pre-implantation embryos after EGA (AUC cutoff > 0.85); (ii) should be significantly upregulated in the majority of analysed naive cultures (3 out of 5).

b. Heatmap showing the comparison of expressional changes as fold-change of naïve vs primed cells with the fold-change of pairwise human and macaque blastocyst stages (ICM, EPI, PE and TE). Row-wise z-score of log2-fold change expression of most variable genes (MVGs) (n=948). The clustered dendrogram represents Spearman's correlation and Euclidian distance. Note the contrasting pattern of gene expression between *in vitro* naïve cultures and human *in vivo* pluripotent states (ICM and EPI) compared with their counter samples. The zoom-in details expressional changes of five genes, whose expression is shifted between macaque and human (e.g. mark pluripotent states in macaque and TE in human). The five genes also represent genes that are upregulated in naïve *in vitro* cultures vs their primed counterparts.

Movie S1

Expression dynamics of L1-ORF1 (red) by immunofluorescence staining during the formation of the blastocyst. Note that L1_HS_ORF1p accumulates in the cytoplasm of Pre-TE and in the blastocoel cells (DAPI, blue). See also Figure S3D.

References

Blakeley, P., Fogarty, N.M., del Valle, I., Wamaitha, S.E., Hu, T.X., Elder, K., Snell, P., Christie, L., Robson, P., and Niakan, K.K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development *142*, 3151-3165.

Braude, P., Bolton, V., and Moore, S. (1988). Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature *332*, 459-461.
Brons, I.G., Smithers, L.E., Trotter, M.W., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A.*, et al.* (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191-195.
Chan, Y.S., Goke, J., Ng, J.H., Lu, X., Gonzales, K.A., Tan, C.P., Tng, W.Q., Hong, Z.Z., Lim, Y.S., and Ng, H.H. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663-675.
Chazaud, C., and Yamanaka, Y. (2016). Lineage specification in the mouse preimplantation embryo. Development *143*, 1063-1074.
Dodsworth, B.T., Flynn, R., and Cowley, S.A. (2015). The Current State of Naive Human Pluripotency. Stem Cells *33*, 3181-3186.
Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E.J., Davila, J., Mall, M., Wong, W.H., Wysocka, J.*, et al.* (2016). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet *48*, 44-52.
Flach, G., Johnson, M.H., Braude, P.R., Taylor, R.A., and Bolton, V.N. (1982). The transition from maternal to embryonic control in the 2-cell mouse embryo. The EMBO journal *1*, 681-686.
Friedli, M., and Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. Annu Rev Cell Dev Biol *31*, 429-451.
Gardner, R.L. (1998). Contributions of blastocyst micromanipulation to the study of mammalian development. Bioessays *20*, 168-180.
Goke, J., Lu, X., Chan, Y.S., Ng, H.H., Ly, L.H., Sachs, F., and Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell *16*, 135-141.
Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y.*, et al.* (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221-225.
Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J.*, et al.* (2014). The DNA methylation landscape of human early embryos. Nature *511*, 606-610.
Hancks, D.C., and Kazazian, H.H., Jr. (2012). Active human retrotransposons: variation and disease. Curr Opin Genet Dev *22*, 191-203.
Hardy, K. (1999). Apoptosis in the human embryo. Rev Reprod *4*, 125-134.
Hu, S., Wilson, K.D., Ghosh, Z., Han, L., Wang, Y., Lan, F., Ransohoff, K.J., Burridge, P., and Wu, J.C. (2013). MicroRNA-302 increases reprogramming efficiency via repression of NR2F2. Stem Cells *31*, 259-268.
Izsvak, Z., Wang, J., Singh, M., Mager, D.L., and Hurst, L.D. (2016). Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? Bioessays *38*, 109-117.
Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. Nature *468*, 811-814.
Knisbacher, B.A., Gerber, D., and Levanon, E.Y. (2016). DNA Editing by APOBECs: A Genomic Preserver and Transformer. Trends Genet *32*, 16-28.
Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S.*, et al.* (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet *42*, 1113-1117.

Lu, X., Sachs, F., Ramsay, L., Jacques, P.E., Goke, J., Bourque, G., and Ng, H.H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nature structural & molecular biology *21*, 423-425.
Macia, A., Widmann, T.J., Heras, S.R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., Munoz-Lopez, M., Rubio, A., Amador-Cubero, S., Blanco-Jimenez, E.*, et al.* (2017). Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Res *27*, 335-348.
Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M.*, et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.
Malki, S., van der Heijden, G.W., O'Donnell, K.A., Martin, S.L., and Bortvin, A. (2014). A role for retrotransposon LINE-1 in fetal oocyte attrition in mice. Developmental cell *29*, 521-533.
Marchetto, M.C.N., Narvaiza, I., Denli, A.M., Benner, C., Lazzarini, T.A., Nathanson, J.L., Paquola, A.C.M., Desai, K.N., Herai, R.H., Weitzman, M.D.*, et al.* (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. Nature *503*, 525-529.
Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007). Which transposable elements are active in the human genome? Trends Genet *23*, 183-191.
Muller, T., Fleischmann, G., Eildermann, K., Matz-Rensing, K., Horn, P.A., Sasaki, E., and Behr, R. (2009). A novel embryonic stem cell line derived from the common marmoset monkey (Callithrix jacchus) exhibiting germ cell-like characteristics. Hum Reprod *24*, 1359-1372.
Nakamura, T., Okamoto, I., Sasaki, K., Yabuta, Y., Iwatani, C., Tsuchiya, H., Seita, Y., Nakamura, S., Yamamoto, T., and Saitou, M. (2016). A developmental coordinate of pluripotency among mice, monkeys and humans. Nature *537*, 57-62.
Ng, S.Y., Johnson, R., and Stanton, L.W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. The EMBO journal *31*, 522-533.
Niakan, K.K., and Eggan, K. (2013). Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. Dev Biol *375*, 54-64.
Niakan, K.K., Han, J., Pedersen, R.A., Simon, C., and Pera, R.A. (2012). Human pre-implantation embryo development. Development *139*, 829-841.
Olson, M.V., and Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet *4*, 20-28.
Pastor, W.A., Chen, D., Liu, W., Kim, R., Sahakyan, A., Lukianchikov, A., Plath, K., Jacobsen, S.E., and Clark, A.T. (2016). Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. Cell Stem Cell *18*, 323-329.
Petropoulos, S., Edsgard, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012-1026.
Ramsay, L., Marchetto, M.C., Caron, M., Chen, S.H., Busche, S., Kwan, T., Pastinen, T., Gage, F.H., and Bourque, G. (2017). Conserved expression of transposon-derived non-coding transcripts in primate stem cells. BMC genomics *18*, 214.
Romer, C., Singh, M., Hurst, L.D., and Izsvak, Z. (2017). How to tame an endogenous retrovirus: HERVH and the evolution of human pluripotency. Curr Opin Virol *25*, 49-58.
Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology *411*, 273-287.
Sahakyan, A., and Plath, K. (2016). Transcriptome Encyclopedia of Early Human Development. Cell *165*, 777-779.
Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat Biotechnol *33*, 495-502.

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. Nature *511*, 611-615.

Suntsova, M., Gogvadze, E.V., Salozhin, S., Gaifullin, N., Eroshkin, F., Dmitriev, S.E., Martynova, N., Kulikov, K., Malakhova, G., Tukhbatova, G.*, et al.* (2013). Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. Proc Natl Acad Sci U S A *110*, 19472-19477.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W.*, et al.* (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254-1269.

Theunissen, T.W., Friedli, M., He, Y., Planet, E., O'Neil, R.C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M.*, et al.* (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. Cell Stem Cell.

Theunissen, T.W., and Jaenisch, R. (2014). Molecular control of induced pluripotency. Cell Stem Cell *14*, 720-734.

van den Hurk, J.A., Meij, I.C., Seleme, M.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S.*, et al.* (2007). L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet *16*, 1587-1592.

Wang, J., Singh, M., Sun, C., Besser, D., Prigione, A., Ivics, Z., Hurst, L.D., and Izsvak, Z. (2016). Isolation and cultivation of naive-like human pluripotent stem cells based on HERVH expression. Nat Protoc *11*, 327-346.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Rasko, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V.*, et al.* (2014a). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature *516*, 405-409.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V.*, et al.* (2014b). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature *516*, 405–409.

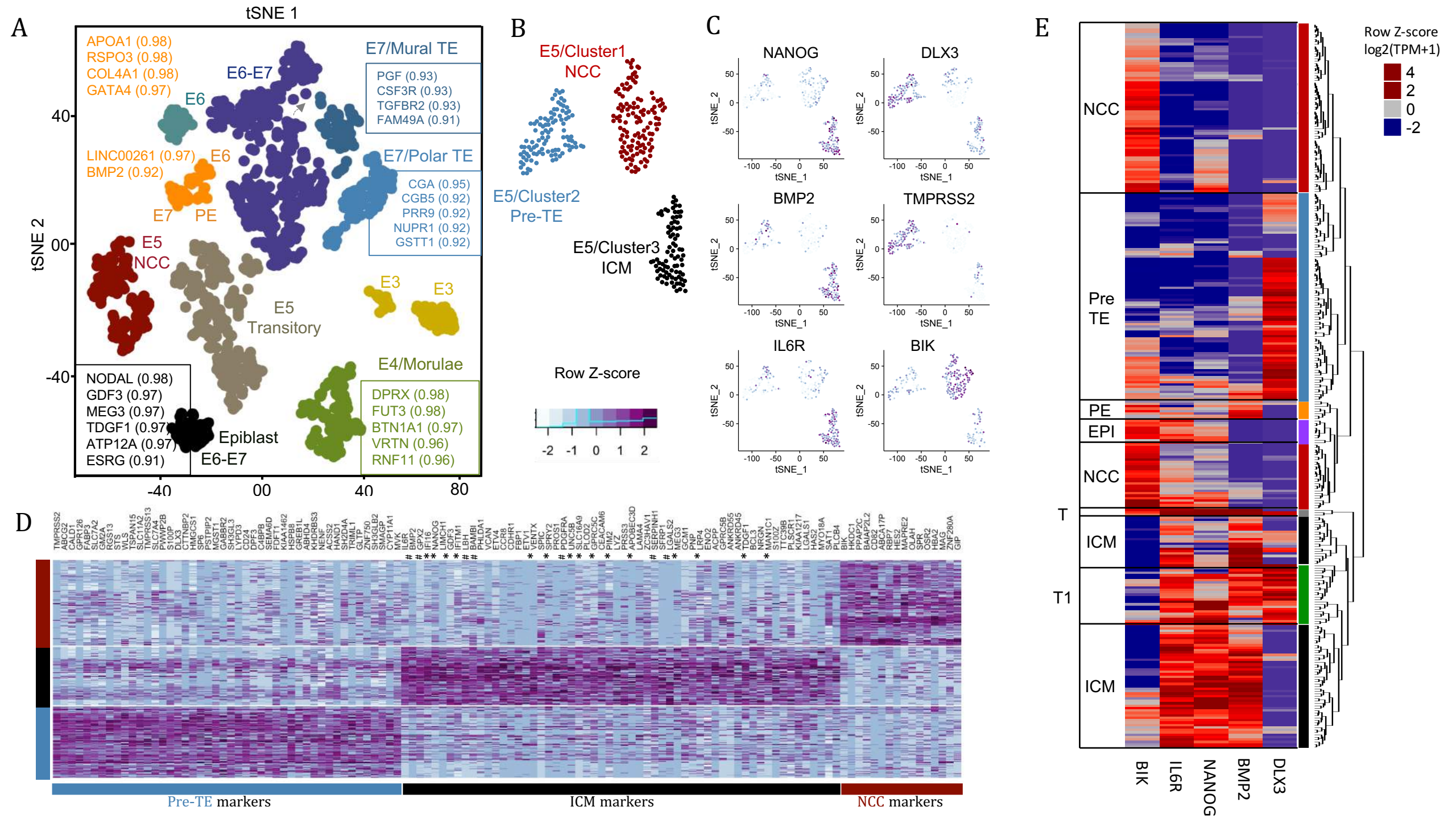
Wunderlich, S., Kircher, M., Vieth, B., Haase, A., Merkert, S., Beier, J., Gohring, G., Glage, S., Schambach, A., Curnow, E.C.*, et al.* (2014). Primate iPS cells as tools for evolutionary analyses. Stem Cell Res *12*, 622-629.

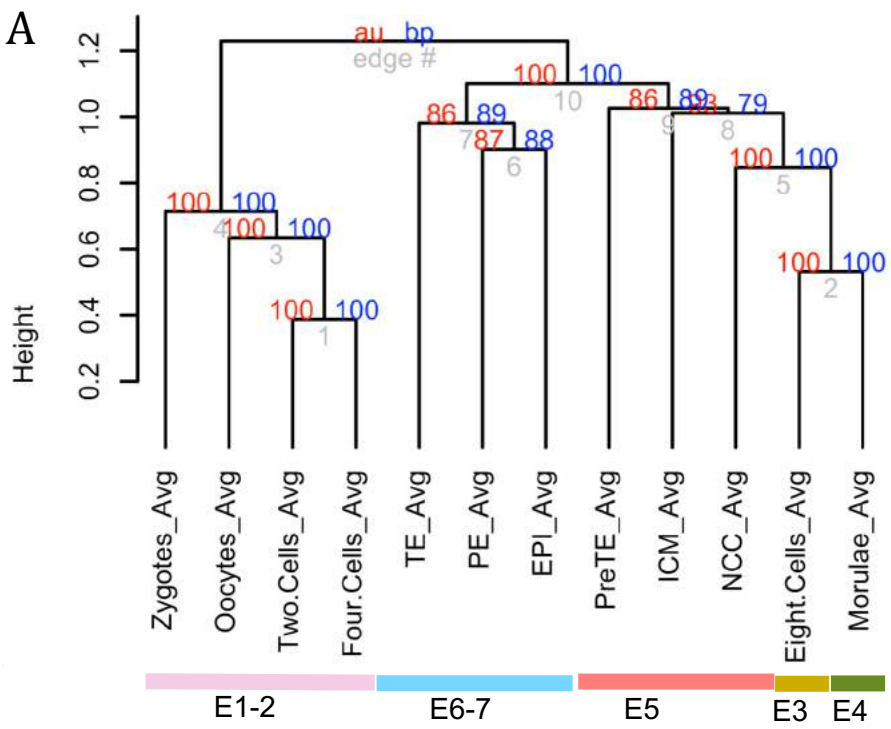
Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J.*, et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology *20*, 1131-1139.

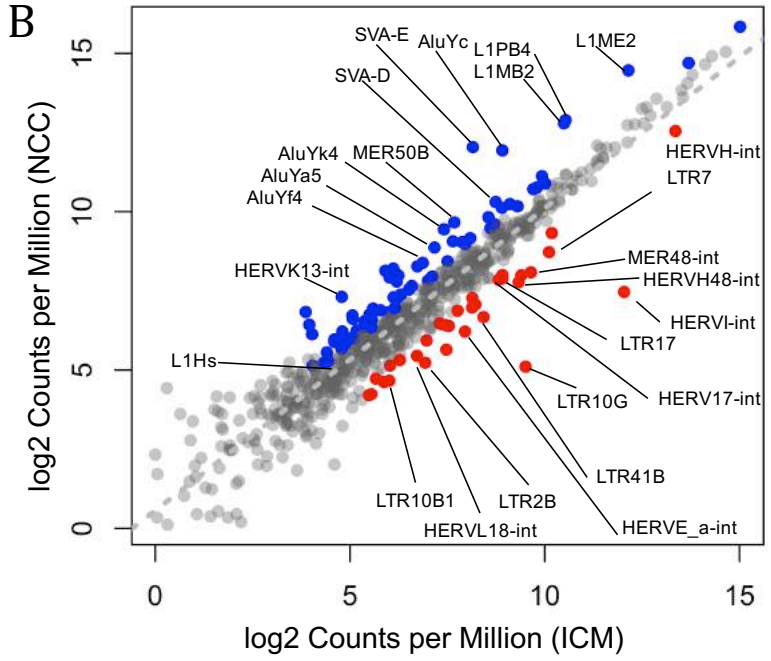
Zhao, M., Ren, C., Yang, H., Feng, X., Jiang, X., Zhu, B., Zhou, W., Wang, L., Zeng, Y., and Yao, K. (2007). Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies HESRG, a novel stem cell gene. Biochem Biophys Res Commun *362*, 916-922.

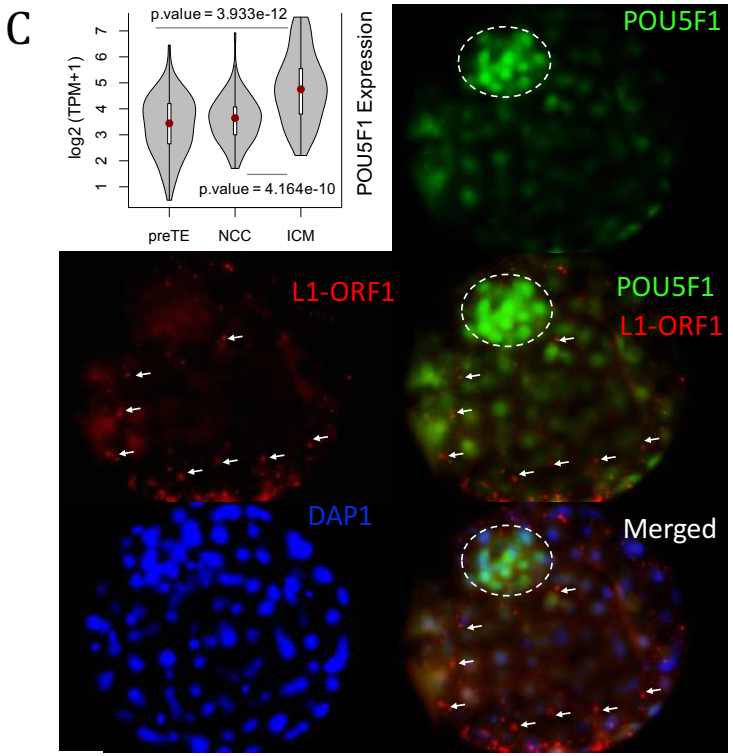
Main Figures

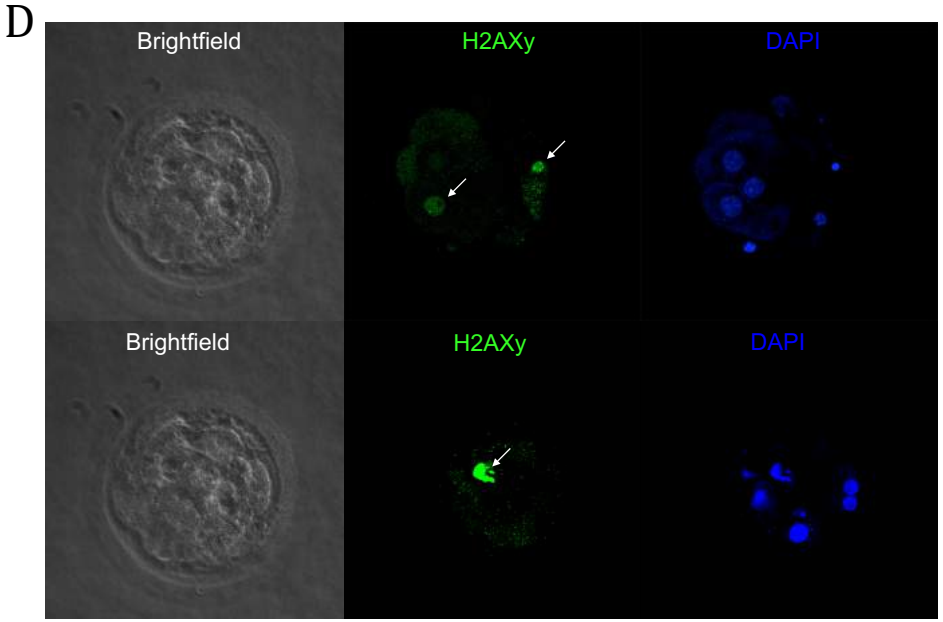
Figure 1 to Figure 6

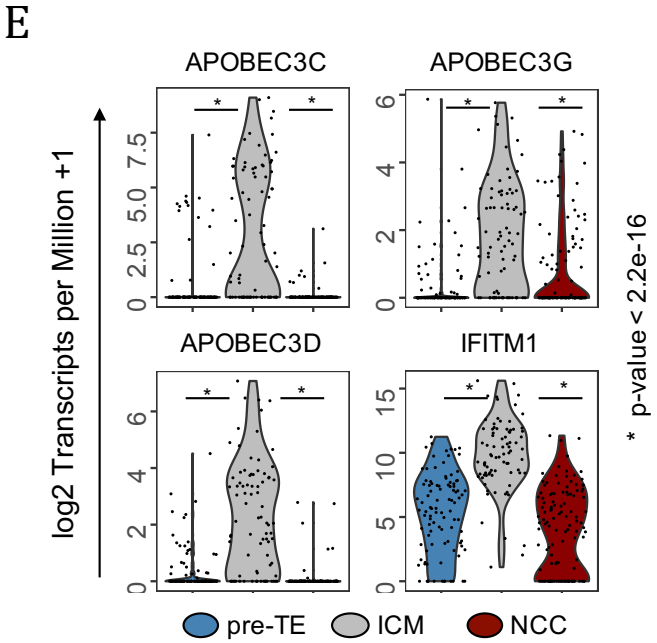

A
tSNE 1
tSNE 2
APOA1 (0.98)
RSPO3 (0.98)
COL4A1 (0.98)
GATA4 (0.97)
E6
E6-E7
E7/Mural TE
PGF (0.93)
CSF3R (0.93)
TGFBR2 (0.93)
FAM49A (0.91)
LINC00261 (0.97)
BMP2 (0.92)
E6
E7/Polar TE
CGA (0.95)
CGB5 (0.92)
PRR9 (0.92)
NUPR1 (0.92)
GSTT1 (0.92)
E7
PE
E5
NCC
E3
E3
E5
Transitory
E4/Morulae
DPRX (0.98)
FUT3 (0.98)
BTN1A1 (0.97)
VRTN (0.96)
RNF11 (0.96)
NODAL (0.98)
GDF3 (0.97)
MEG3 (0.97)
TDGF1 (0.97)
ATP12A (0.97)
ESRG (0.91)
Epiblast
E6-E7
B
E5/Cluster1
NCC
E5/Cluster2
Pre-TE
E5/Cluster3
ICM
Row Z-score
C
NANOG
DLX3
BMP2
TMPRSS2
IL6R
BIK
tSNE_1
tSNE_2
E
Row Z-score
log2(TPM+1)
NCC
Pre
TE
PE
EPI
NCC
ICM
ICM
T
T1
BIK
IL6R
NANOG
BMP2
DLX3
D
Pre-TE markers
ICM markers
NCC markers

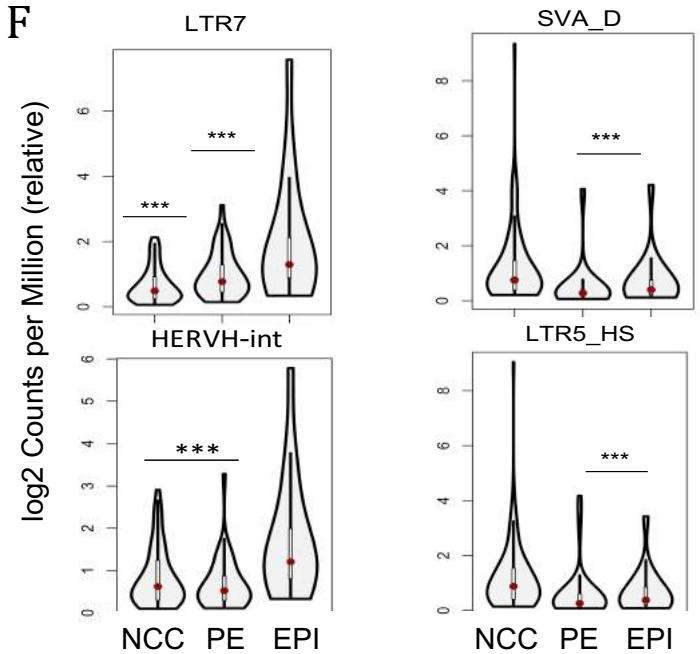

A
au bp
edge #
Height
Zygotes_Avg
Oocytes_Avg
Two.Cells_Avg
Four.Cells_Avg
TE_Avg
PE_Avg
EPI_Avg
PreTE_Avg
ICM_Avg
NCC_Avg
Eight.Cells_Avg
Morulae_Avg
E1-2
E6-7
E5
E3
E4

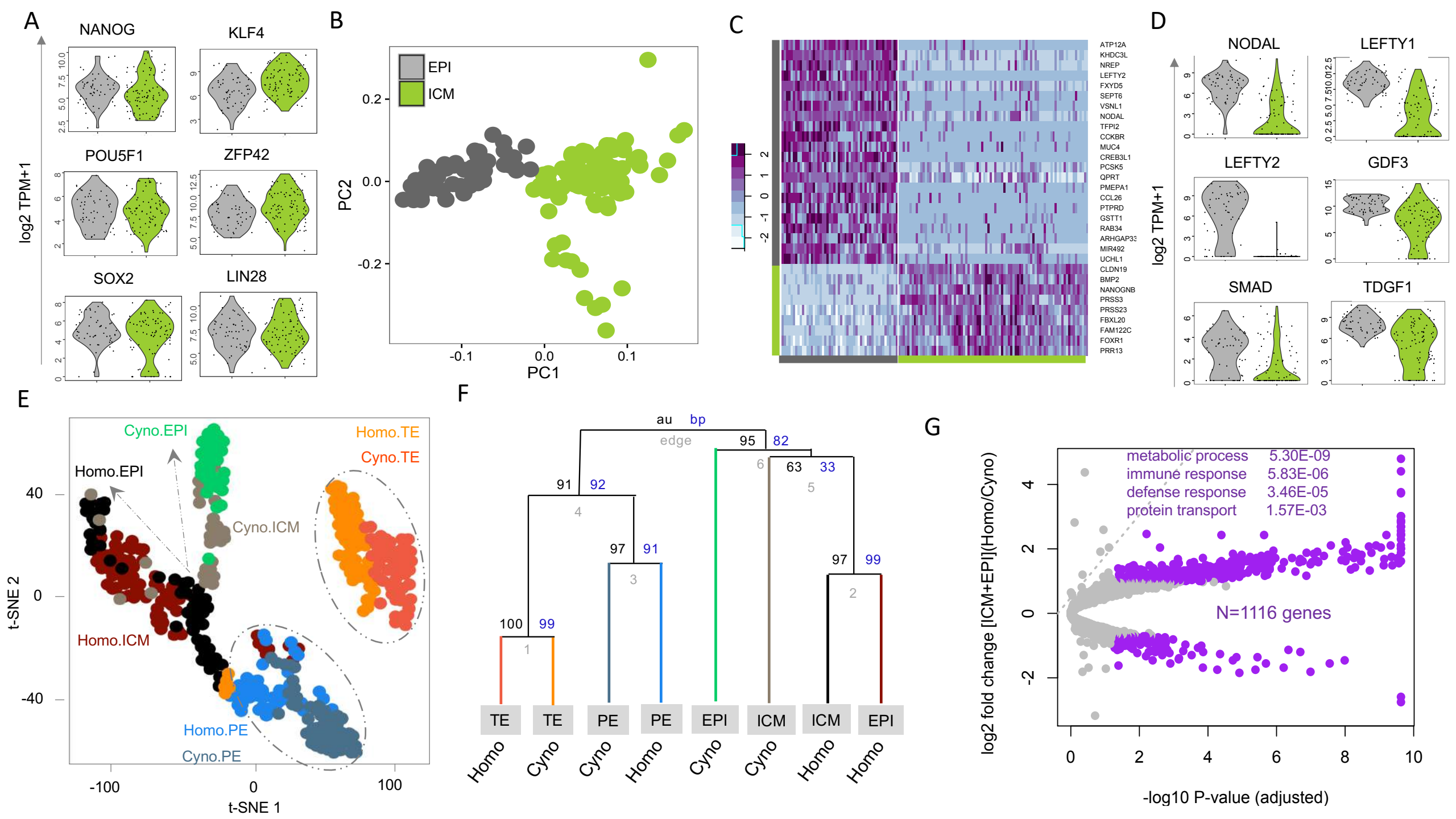

B
SVA-E
AluYc
L1PB4
L1ME2
SVA-D
L1MB2
AluYk4
MER50B
AluYa5
AluYf4
HERVH-int
LTR7
HERVK13-int
MER48-int
HERVH48-int
HERVI-int
LTR17
L1Hs
LTR10G
HERV17-int
LTR41B
LTR10B1
LTR2B
HERVL18-int
HERVE_a-int
log2 Counts per Million (NCC)
log2 Counts per Million (ICM)

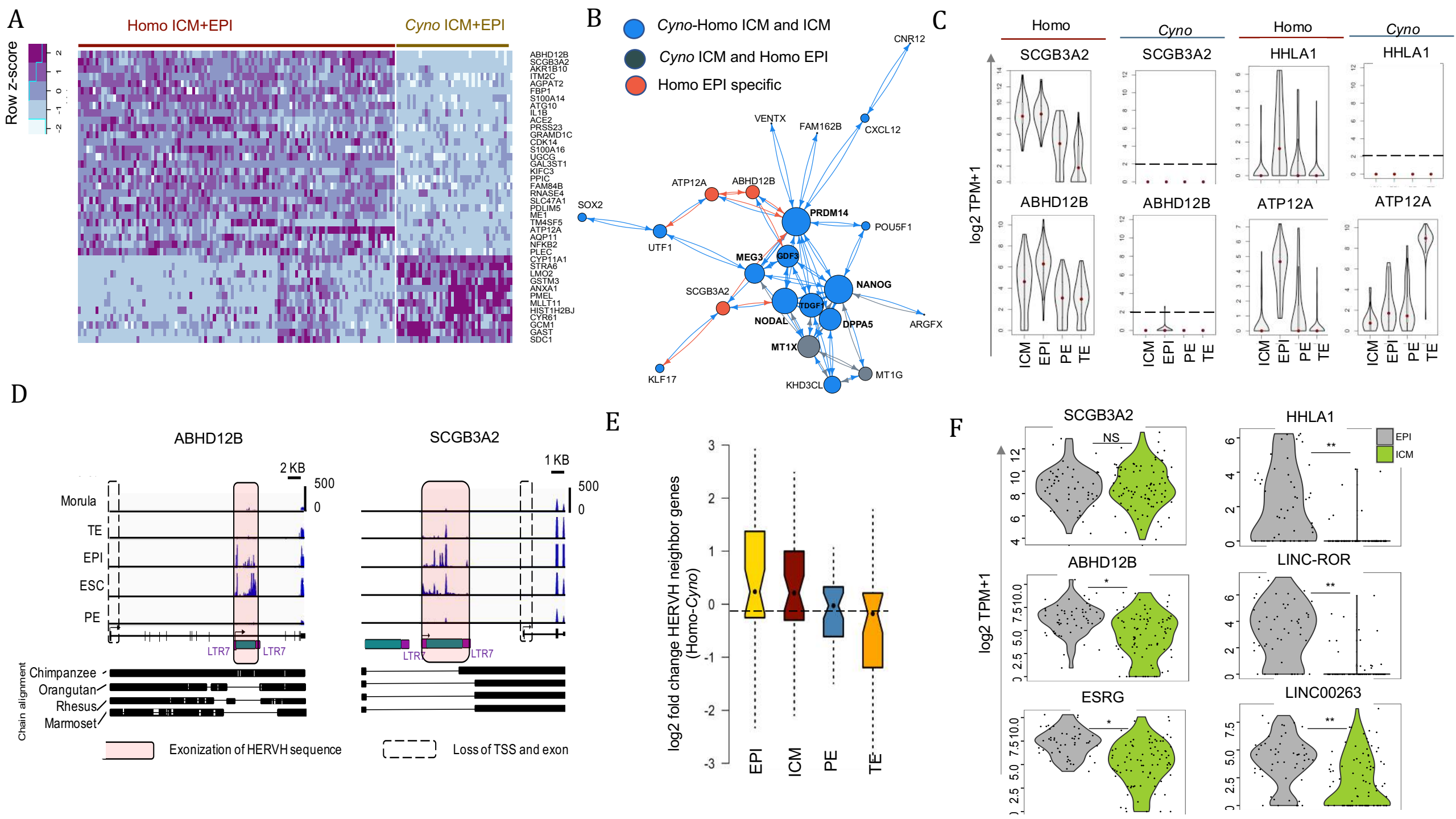

C
p.value = 3.933e-12
p.value = 4.164e-10
log2 (TPM+1)
POU5F1 Expression
preTE
NCC
ICM
POU5F1
L1-ORF1
POU5F1
L1-ORF1
DAP1
Merged

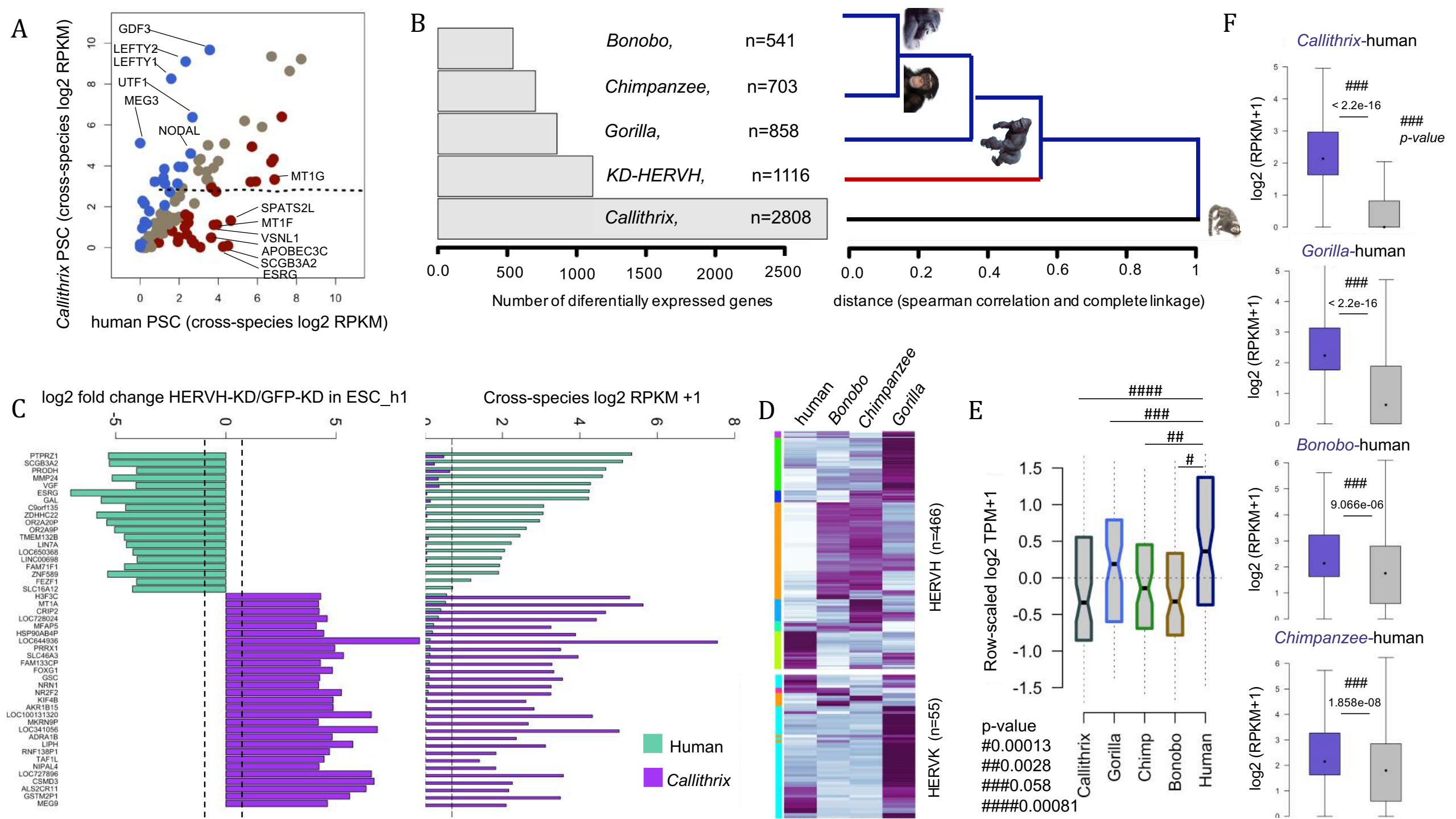

D
Brightfield
H2AXy
DAPI
Brightfield
H2AXy
DAPI

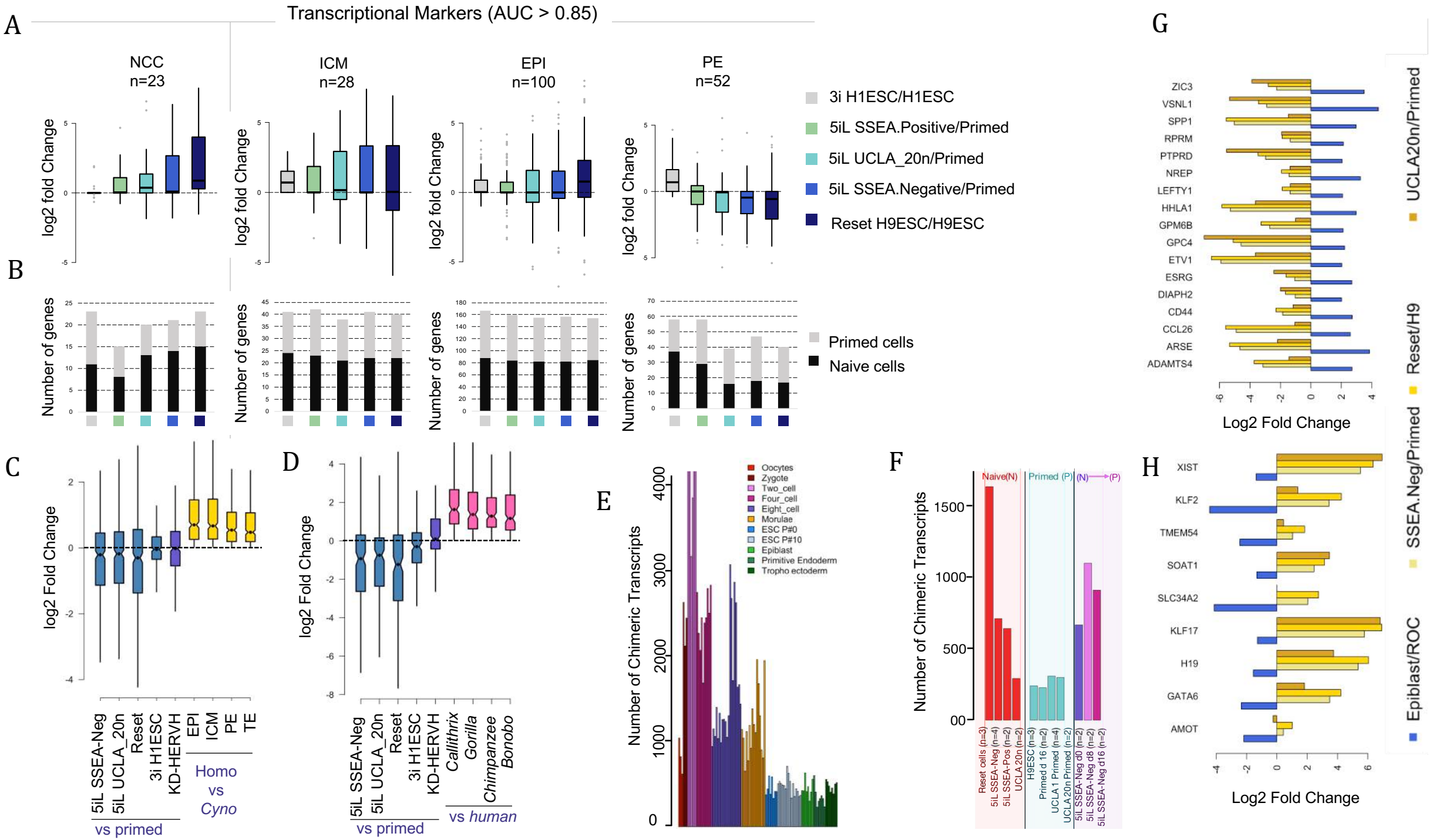

E
APOBEC3C
APOBEC3G
APOBEC3D
IFITM1
log2 Transcripts per Million +1
* p-value < 2.2e-16
pre-TE
ICM
NCC


F
LTR7
SVA_D
HERVH-int
LTR5_HS
log2 Counts per Million (relative)
NCC
PE
EPI
NCC
PE
EPI


A
NANOG
KLF4
POU5F1
ZFP42
SOX2
LIN28
log2 TPM+1
B
EPI
ICM
PC2
PC1
C
ATP12A
KHDC3L
NREP
LEFTY2
FXYD5
SEPT6
VSNL1
NODAL
TFPI2
CCKBR
MUC4
CREB3L1
PCSK5
QPRT
PMEPA1
CCL26
PTPRD
GSTT1
RAB34
ARHGAP33
MIR492
UCHL1
CLDN19
BMP2
NANOGNB
PRSS3
PRSS23
FBXL20
FAM122C
FOXR1
PRR13
D
NODAL
LEFTY1
LEFTY2
GDF3
SMAD
TDGF1
log2 TPM+1
E
Cyno.EPI
Homo.TE
Cyno.TE
Homo.EPI
Cyno.ICM
Homo.ICM
Homo.PE
Cyno.PE
t-SNE 2
t-SNE 1
F
au
bp
edge
TE
TE
PE
PE
EPI
ICM
ICM
EPI
Homo
Cyno
Cyno
Homo
Cyno
Cyno
Homo
Homo
G
metabolic process 5.30E-09
immune response 5.83E-06
defense response 3.46E-05
protein transport 1.57E-03
N=1116 genes
log2 fold change [ICM+EPI](Homo/Cyno)
-log10 P-value (adjusted)

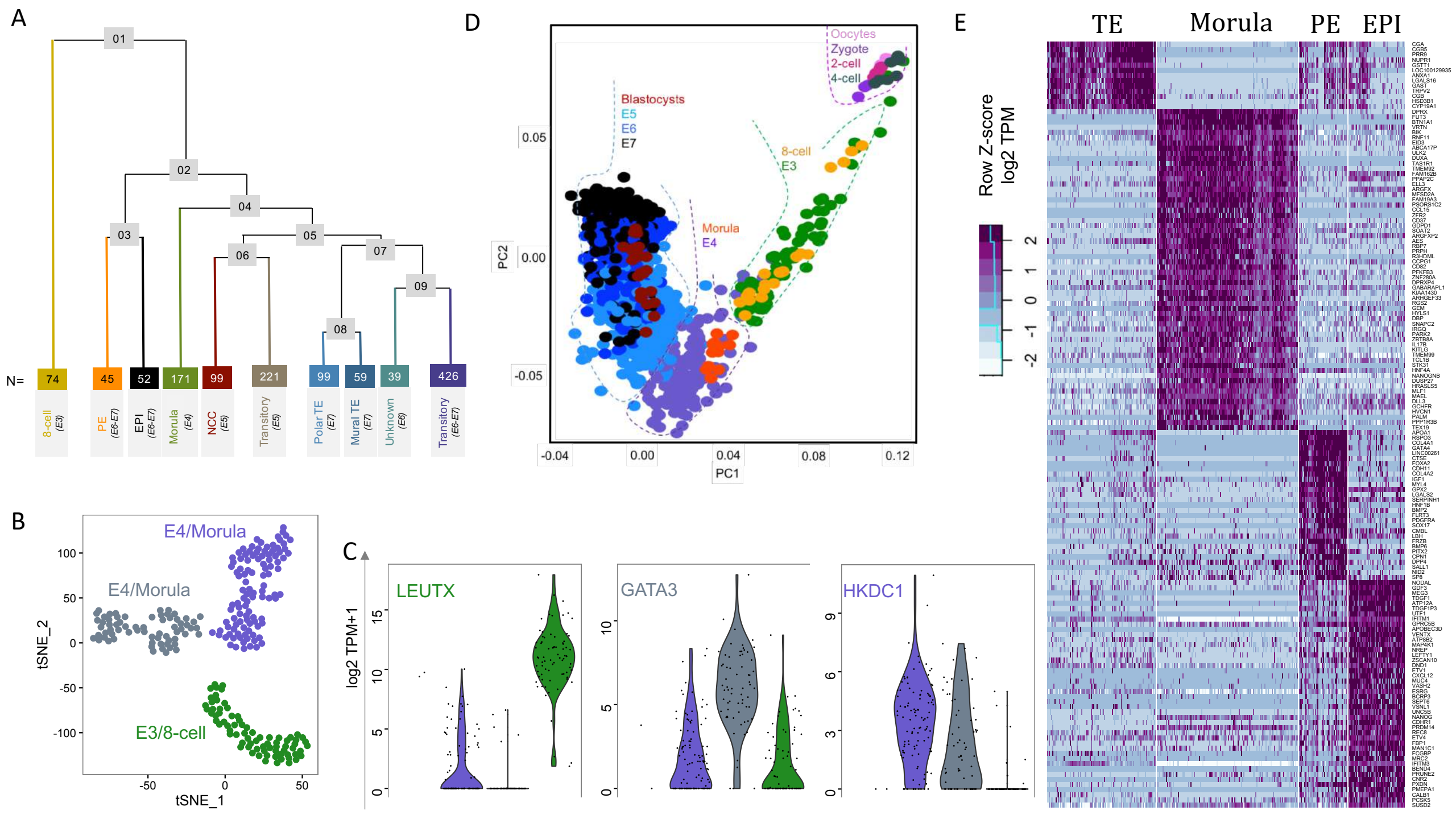

A
Homo ICM+EPI
Cyno ICM+EPI
Row z-score
ABHD12B
SCGB3A2
AKR1B10
ITM2C
AGPAT2
FBP1
S100A14
ATG10
IL1B
ACE2
PRSS23
GRAMD1C
CDK14
S100A16
UGCG
GAL3ST1
KIFC3
PPIC
FAM84B
RNASE4
SLC47A1
PDLIM5
ME1
TM4SF5
ATP12A
AQP11
NFKB2
PLEC
CYP11A1
STRA6
LMO2
GSTM3
ANXA1
PMEL
MLLT11
HIST1H2BJ
CYR61
GCM1
GAST
SDC1
B
Cyno-Homo ICM and ICM
Cyno ICM and Homo EPI
Homo EPI specific
CNR12
VENTX
FAM162B
CXCL12
ATP12A
ABHD12B
SOX2
PRDM14
POU5F1
UTF1
MEG3
GDF3
NANOG
SCGB3A2
TDGF1
NODAL
DPPA5
ARGFX
MT1X
MT1G
KLF17
KHD3CL
C
Homo
Cyno
Homo
Cyno
SCGB3A2
SCGB3A2
HHLA1
HHLA1
ABHD12B
ABHD12B
ATP12A
ATP12A
log2 TPM+1
ICM
EPI
PE
TE
D
ABHD12B
SCGB3A2
2 KB
1 KB
500
0
Morula
TE
EPI
ESC
PE
LTR7
LTR7
Chimpanzee
Orangutan
Rhesus
Marmoset
Chain alignment
Exonization of HERVH sequence
Loss of TSS and exon
E
log2 fold change HERVH neighbor genes (Homo-Cyno)
EPI
ICM
PE
TE
F
SCGB3A2
HHLA1
ABHD12B
LINC-ROR
ESRG
LINC00263
NS
EPI
ICM
log2 TPM+1

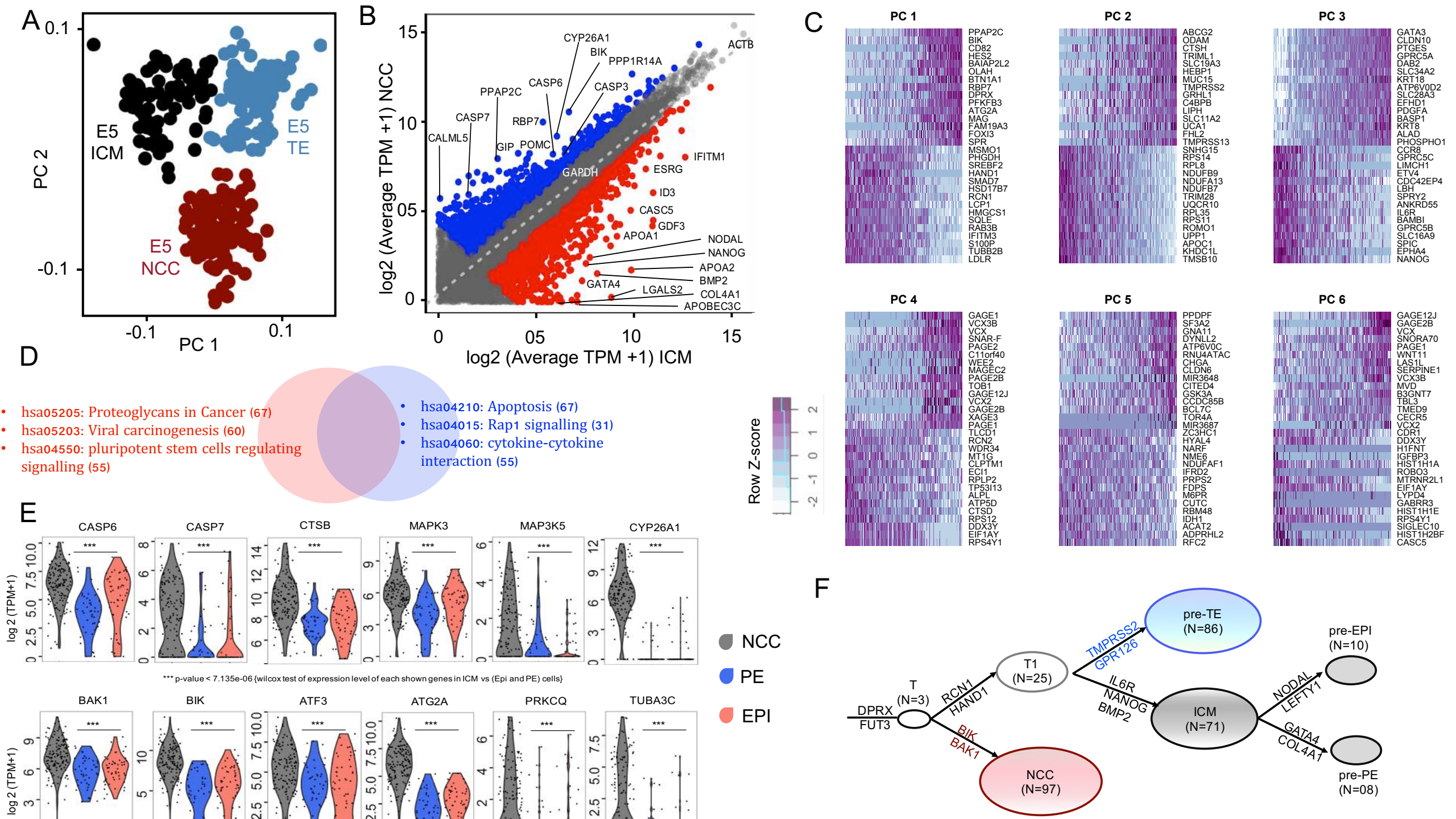

A
GDF3
LEFTY2
LEFTY1
UTF1
MEG3
NODAL
MT1G
SPATS2L
MT1F
VSNL1
APOBEC3C
SCGB3A2
ESRG
Callithrix PSC (cross-species log2 RPKM)
human PSC (cross-species log2 RPKM)
B
Bonobo, n=541
Chimpanzee, n=703
Gorilla, n=858
KD-HERVH, n=1116
Callithrix, n=2808
Number of diferentially expressed genes
distance (spearman correlation and complete linkage)
C
log2 fold change HERVH-KD/GFP-KD in ESC_h1
Cross-species log2 RPKM +1
Human
Callithrix
D
human
Bonobo
Chimpanzee
Gorilla
HERVH (n=466)
HERVK (n=55)
E
Row-scaled log2 TPM+1
Callithrix
Gorilla
Chimp
Bonobo
Human
p-value
#0.00013
##0.0028
###0.058
####0.00081
F
Callithrix-human
< 2.2e-16
p-value
Gorilla-human
< 2.2e-16
Bonobo-human
9.066e-06
Chimpanzee-human
1.858e-08
log2 (RPKM+1)

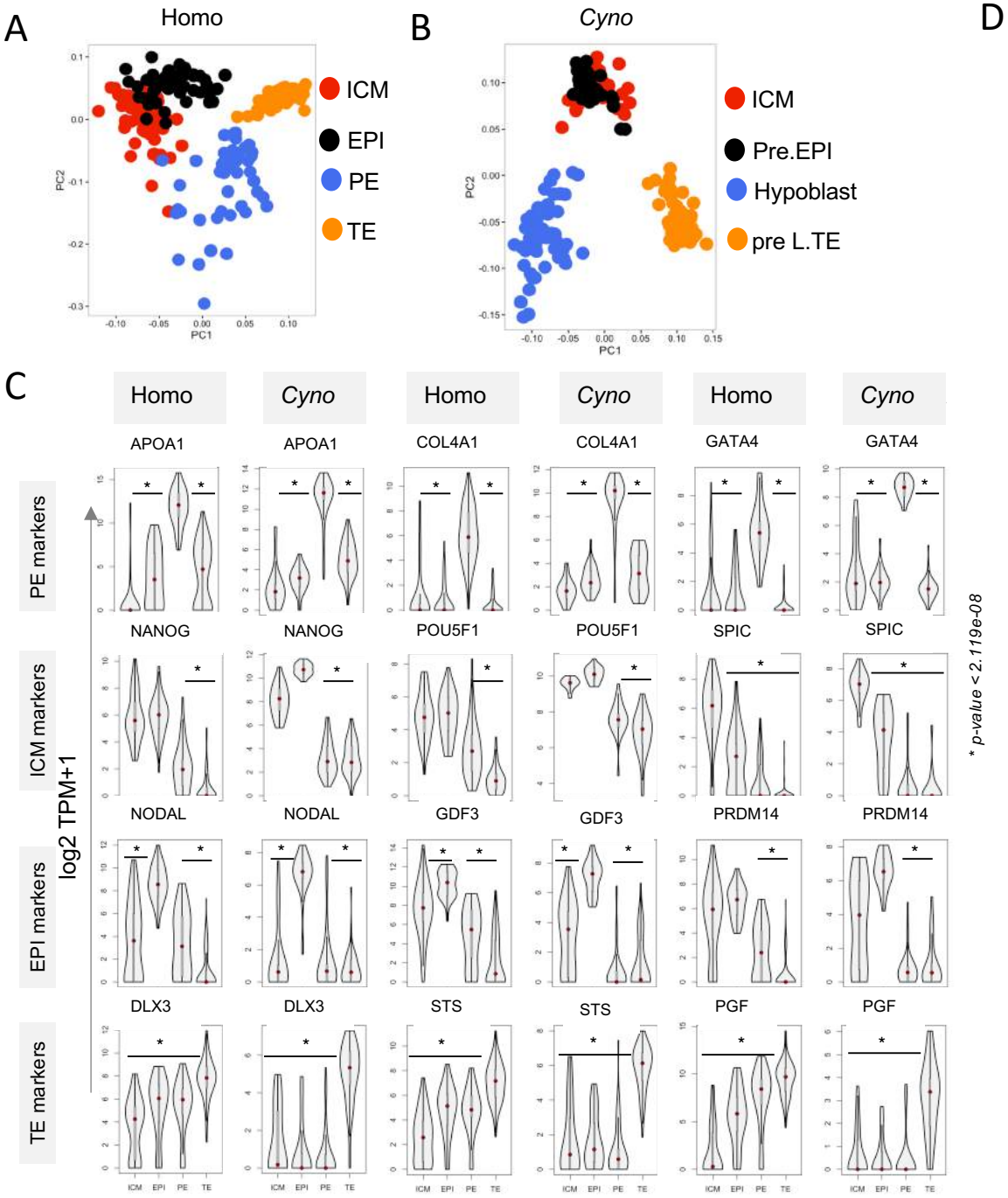

A
Transcriptional Markers (AUC > 0.85)
NCC
n=23
ICM
n=28
EPI
n=100
PE
n=52
log2 fold Change
3i H1ESC/H1ESC
5iL SSEA.Positive/Primed
5iL UCLA_20n/Primed
5iL SSEA.Negative/Primed
Reset H9ESC/H9ESC
B
Number of genes
Primed cells
Naive cells
C
log2 Fold Change
5iL SSEA-Neg
5iL UCLA_20n
Reset
3i H1ESC
KD-HERVH
EPI
ICM
PE
TE
vs primed
Homo
vs
Cyno
D
log2 Fold Change
5iL SSEA-Neg
5iL UCLA_20n
Reset
3i H1ESC
KD-HERVH
Callithrix
Gorilla
Chimpanzee
Bonobo
vs primed
vs human
E
Number of Chimeric Transcripts
Oocytes
Zygote
Two_cell
Four_cell
Eight_cell
Morulae
ESC P#0
ESC P#10
Epiblast
Primitive Endoderm
Tropho ectoderm
F
Number of Chimeric Transcripts
Naive(N)
Primed (P)
(N)→(P)
Reset cells (n=3)
5iL SSEA-Neg (n=4)
5iL SSEA-Pos (n=2)
UCLA 20n (n=2)
H9ESC (n=3)
Primed d 16 (n=2)
UCLA 1 Primed (n=4)
UCLA 20n Primed (n=2)
5iL SSEA-Neg d0 (n=2)
5iL SSEA-Neg d8 (n=2)
5iL SSEA-Neg d16 (n=2)
G
ZIC3
VSNL1
SPP1
RPRM
PTPRD
NREP
LEFTY1
HHLA1
GPM6B
GPC4
ETV1
ESRG
DIAPH2
CD44
CCL26
ARSE
ADAMTS4
Log2 Fold Change
UCLA20n/Primed
Reset/H9
SSEA.Neg/Primed
Epiblast/ROC
H
XIST
KLF2
TMEM54
SOAT1
SLC34A2
KLF17
H19
GATA6
AMOT
Log2 Fold Change

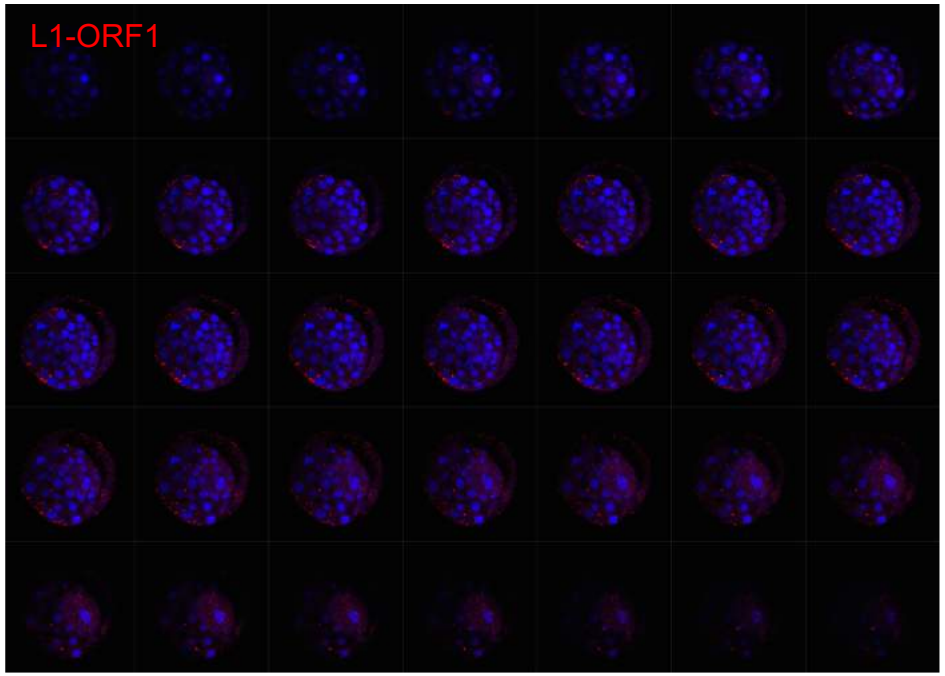
Supplementary Figures

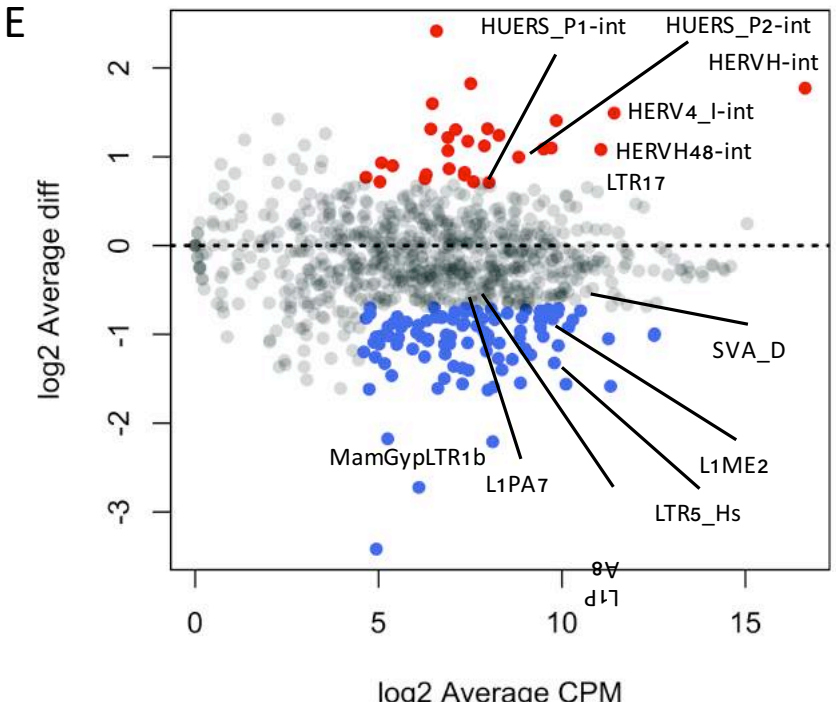
S. Figure 1 to S. Figure 6


A
01
02
03
04
05
06
07
08
09
N=
74
45
52
171
99
221
99
59
39
426
8-cell (E3)
PE (E6-E7)
EPI (E6-E7)
Morula (E4)
NCC (E5)
Transitory (E5)
Polar TE (E7)
Mural TE (E7)
Unknown (E6)
Transitory (E6-E7)
B
E4/Morula
E4/Morula
E3/8-cell
tSNE_2
tSNE_1
C
LEUTX
GATA3
HKDC1
log2 TPM+1
D
Oocytes
Zygote
2-cell
4-cell
Blastocysts
E5
E6
E7
8-cell
E3
Morula
E4
PC2
PC1
E
TE
Morula
PE
EPI
Row Z-score
log2 TPM

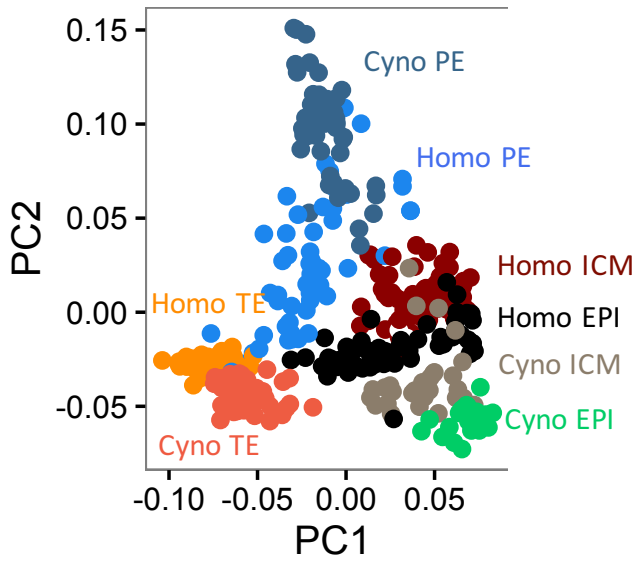

A
E5 ICM
E5 TE
E5 NCC
PC 1
PC 2
B
log2 (Average TPM +1) NCC
log2 (Average TPM +1) ICM
CYP26A1
BIK
PPP1R14A
ACTB
CASP6
CASP3
PPAP2C
CASP7
RBP7
CALML5
GIP
POMC
GAPDH
IFITM1
ESRG
ID3
CASC5
GDF3
APOA1
NODAL
NANOG
APOA2
BMP2
GATA4
LGALS2
COL4A1
APOBEC3C
C
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
Row Z-score
D
hsa05205: Proteoglycans in Cancer (67)
hsa05203: Viral carcinogenesis (60)
hsa04550: pluripotent stem cells regulating signalling (55)
hsa04210: Apoptosis (67)
hsa04015: Rap1 signalling (31)
hsa04060: cytokine-cytokine interaction (55)
E
CASP6
CASP7
CTSB
MAPK3
MAP3K5
CYP26A1
BAK1
BIK
ATF3
ATG2A
PRKCQ
TUBA3C
log 2 (TPM+1)
*** p-value < 7.135e-06 (wilcox test of expression level of each shown genes in ICM vs (Epi and PE) cells)
NCC
PE
EPI
F
DPRX
FUT3
T (N=3)
RCN1
HAND1
T1 (N=25)
BIK
BAK1
NCC (N=97)
TMPRSS2
GPR126
pre-TE (N=86)
IL6R
NANOG
BMP2
ICM (N=71)
NODAL
LEFTY1
pre-EPI (N=10)
GATA4
COL4A1
pre-PE (N=08)


A
Homo
ICM
EPI
PE
TE
PC1
PC2
B
Cyno
ICM
Pre.EPI
Hypoblast
pre L.TE
PC1
PC2
C
Homo
Cyno
APOA1
COL4A1
GATA4
NANOG
POU5F1
SPIC
NODAL
GDF3
PRDM14
DLX3
STS
PGF
PE markers
ICM markers
EPI markers
TE markers
log2 TPM+1
ICM EPI PE TE
* p-value < 2.119e-08

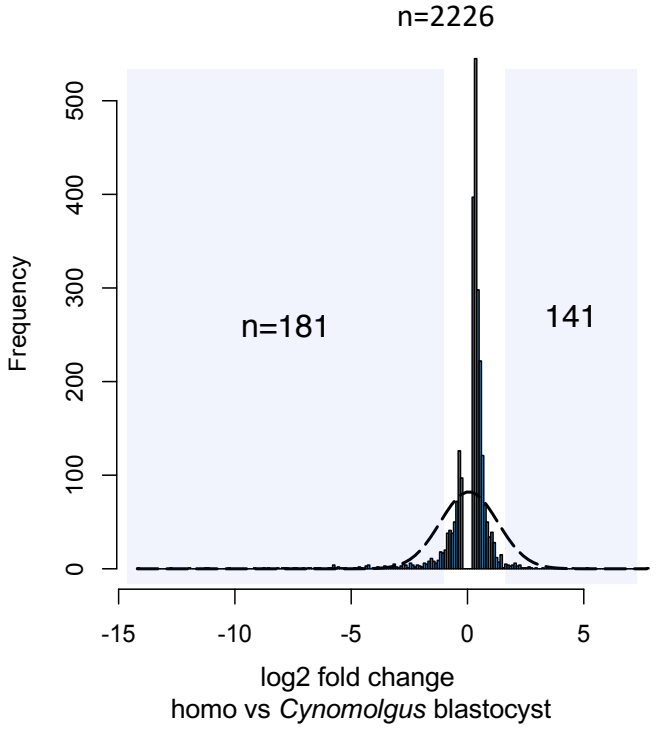

D
L1-ORF1


E
HUERS_P1-int
HUERS_P2-int
HERVH-int
HERV4_I-int
HERVH48-int
LTR17
SVA_D
L1ME2
LTR5_Hs
L1PA7
MamGypLTR1b
L1P
A8
log2 Average diff
log2 Average CPM

A



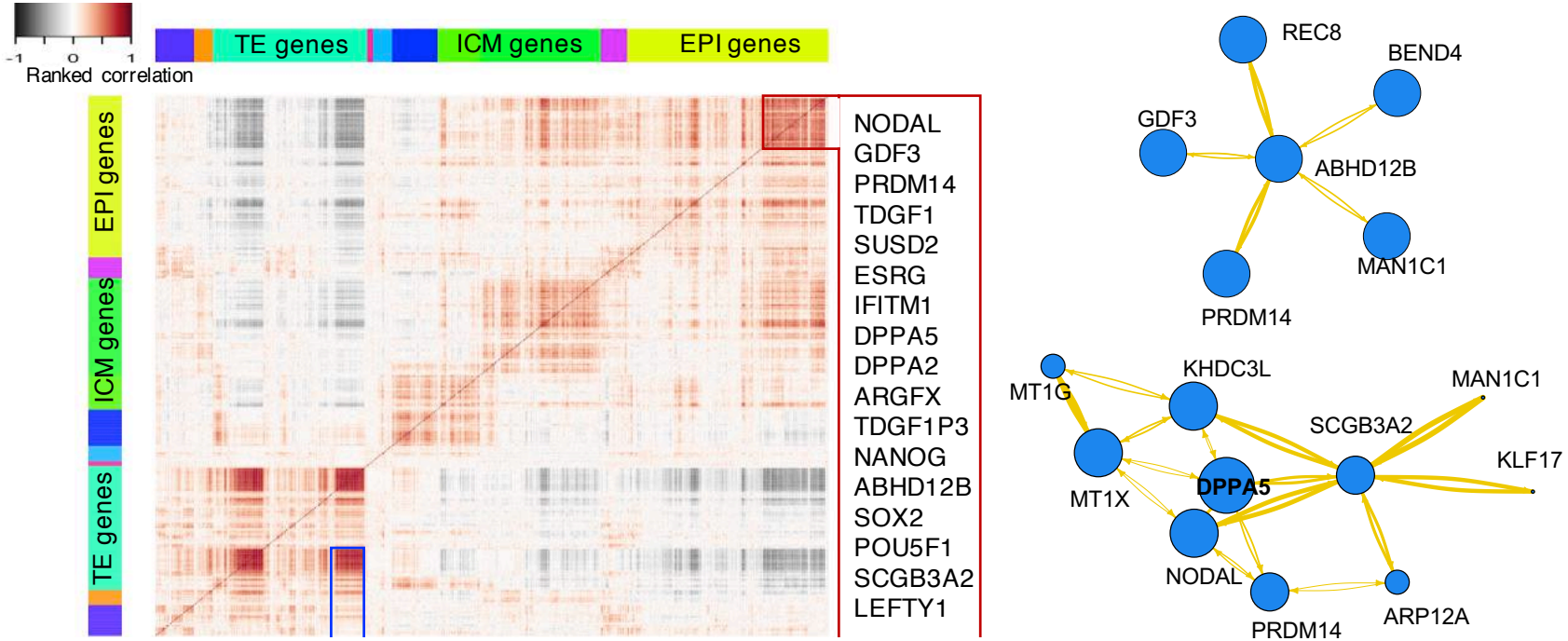
B



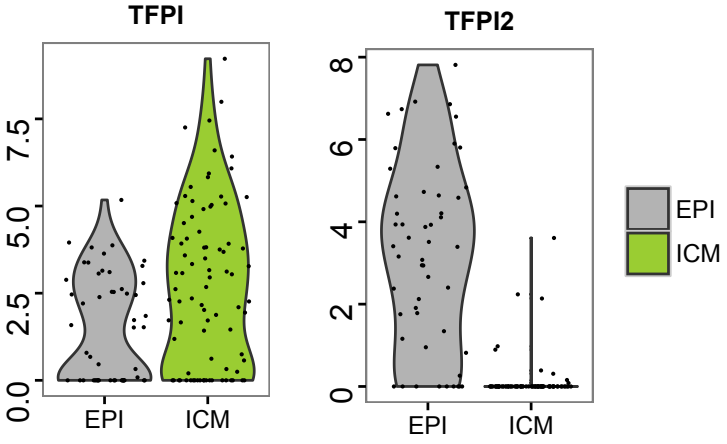
C

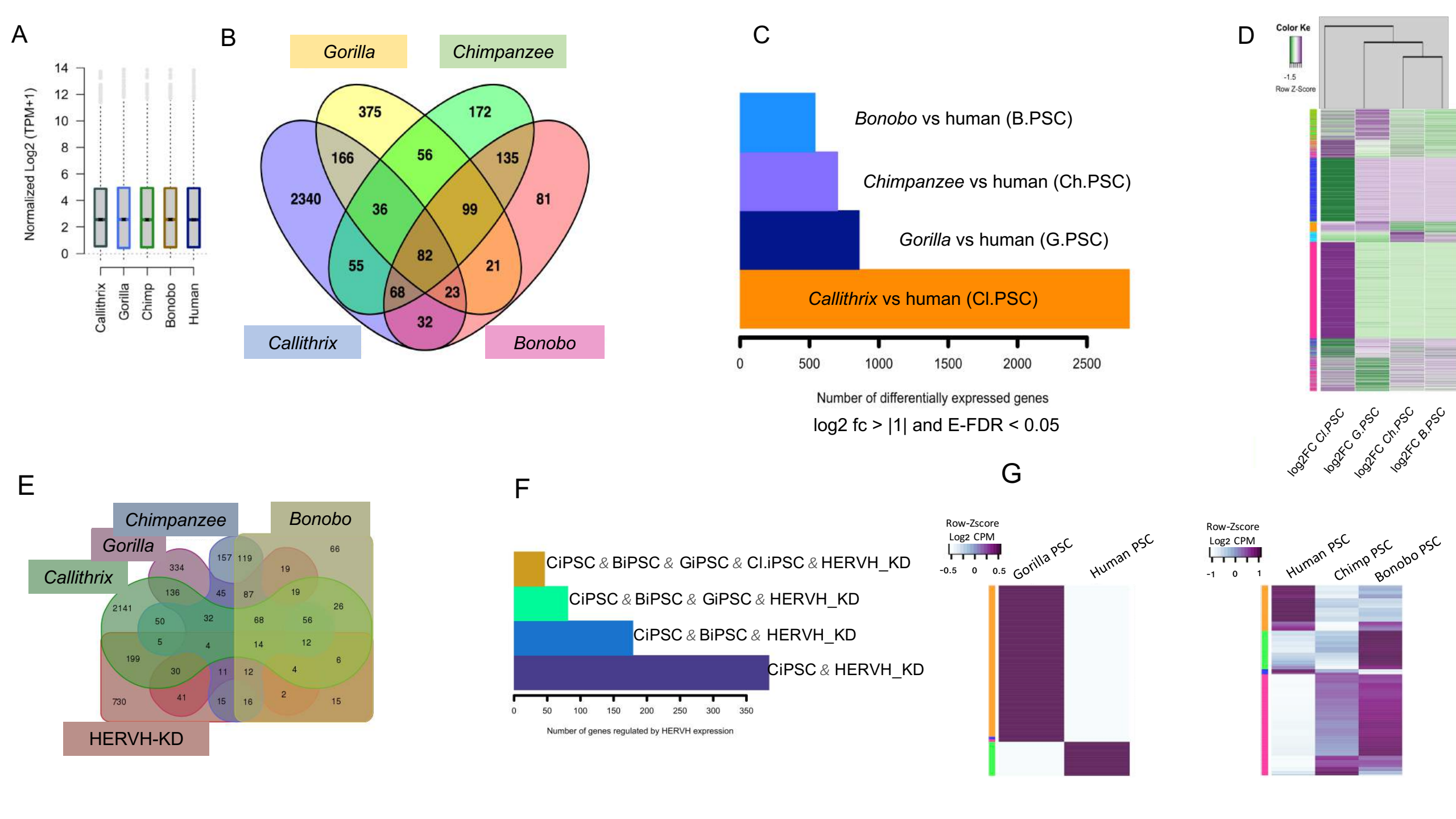
GO ID	GO Term	P-value
GO:0048856	anatomical structure development	5.55E-06
GO:0006954	inflammatory response	1.56E-5
GO:0009725	response to hormone	3.45E-5
GO:0044273	sulfur compound catabolic process	9.35E-5
GO:0006952	defense response	1.07E-4
GO:0060191	regulation of lipase activity	3.72E-4
GO:0044421	extracellular region part	5.42E-4
GO:0050777	negative regulation of immune response	4.08E-4
GO:0032652	regulation of interleukin-1 production	6.42E-4
GO:0030154	cell differentiation	9.14E-4

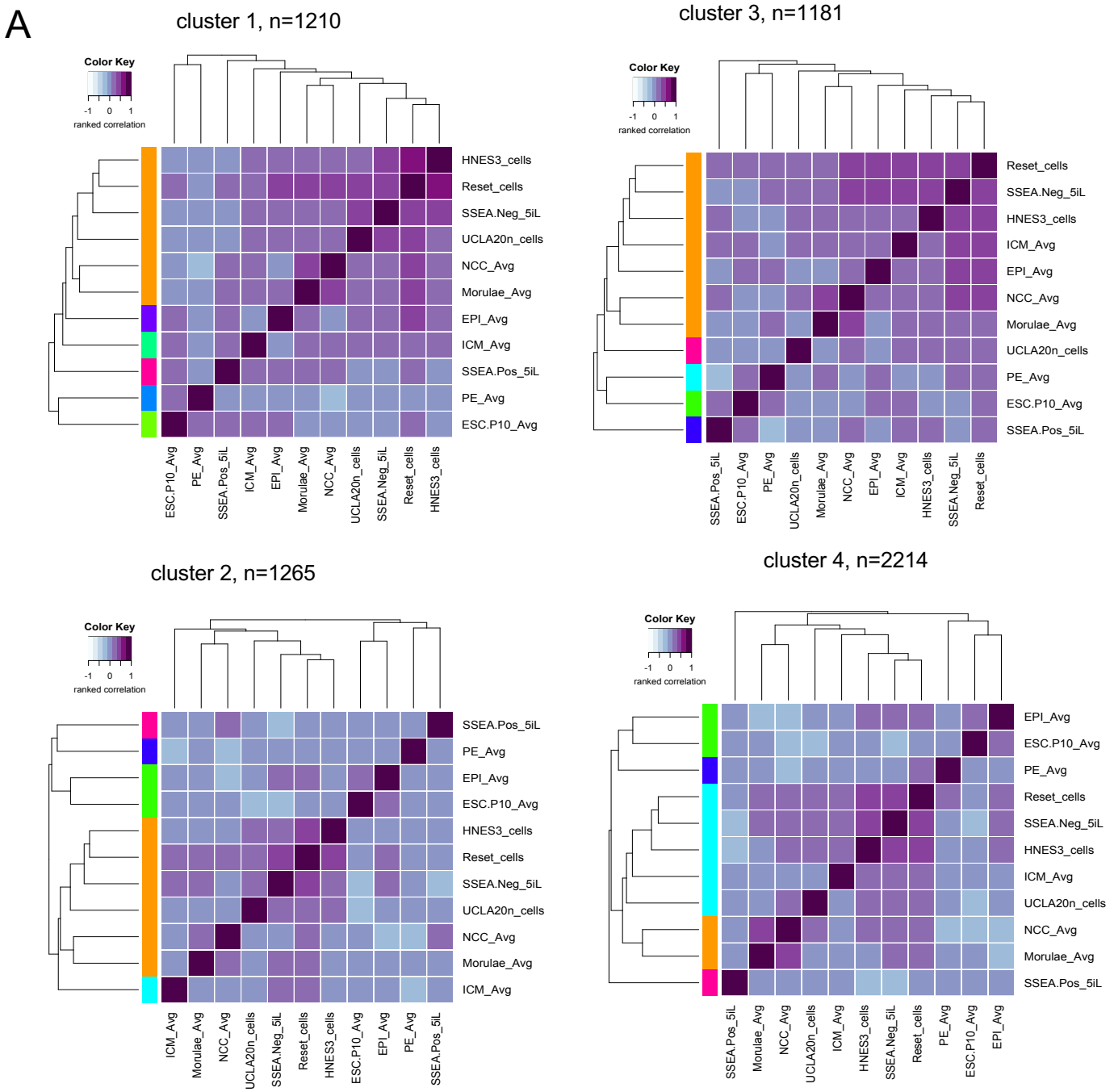
D

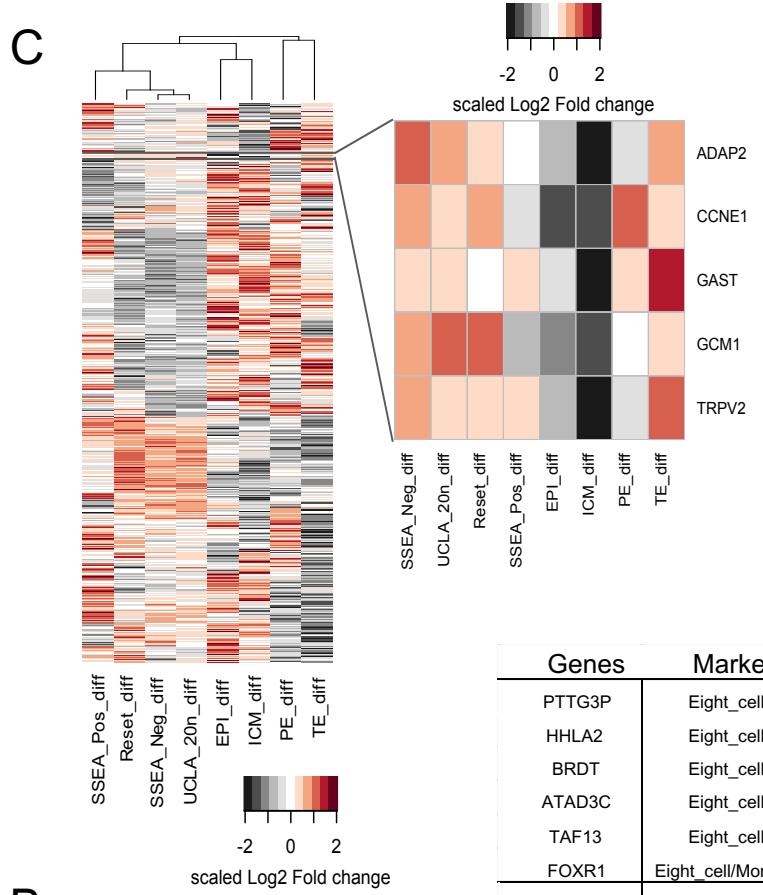


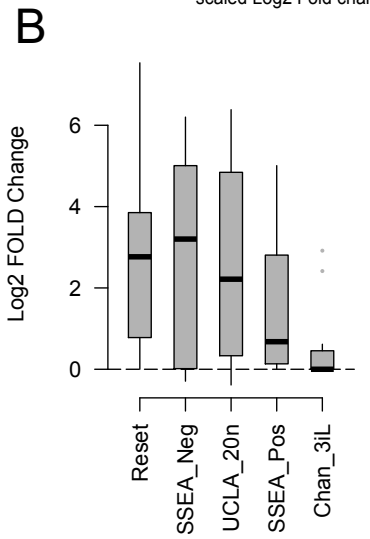
E




A
Normalized Log2 (TPM+1)
14
12
10
8
6
4
2
0
Callithrix
Gorilla
Chimp
Bonobo
Human
B
Gorilla
Chimpanzee
Callithrix
Bonobo
375
172
166
56
135
2340
36
99
81
82
55
21
68
23
32
C
Bonobo vs human (B.PSC)
Chimpanzee vs human (Ch.PSC)
Gorilla vs human (G.PSC)
Callithrix vs human (Cl.PSC)
0
500
1000
1500
2000
2500
Number of differentially expressed genes
log2 fc > |1| and E-FDR < 0.05
D
Color Ke
-1.5
Row Z-Score
log2FC Cl.PSC
log2FC G.PSC
log2FC Ch.PSC
log2FC B.PSC
E
Chimpanzee
Bonobo
Gorilla
Callithrix
HERVH-KD
F
CiPSC & BiPSC & GiPSC & Cl.iPSC & HERVH_KD
CiPSC & BiPSC & GiPSC & HERVH_KD
CiPSC & BiPSC & HERVH_KD
CiPSC & HERVH_KD
0
50
100
150
200
250
300
350
Number of genes regulated by HERVH expression
G
Row-Zscore
Log2 CPM
-0.5
0
0.5
Gorilla PSC
Human PSC
Row-Zscore
Log2 CPM
-1
0
1
Human PSC
Chimp PSC
Bonobo PSC







Genes	Markers	AUC %
PTTG3P	Eight_cell	95
HHLA2	Eight_cell	95
BRDT	Eight_cell	94
ATAD3C	Eight_cell	92
TAF13	Eight_cell	90
FOXR1	Eight_cell/Morulae	96/88
IER5	Morulae	89
TRIM60	Morulae	87
GINS3	Morulae	87
YPEL2	Morulae	86
FAM151A	Morulae/NCC	90/86
DEFB122	NCC	87
OLAH	NCC	87
PLLP	NCC	88
PPP1R14A	NCC	85
IL6R	Transitory	88
GPX2	PE	92
GYPC	PE	88
FBP1	EPI	92
SUSD2	EPI	90
LIM2	EPI	86
TSPAN18	EPI	86