Stem Cell Research 37 (2019) 101456

Contents lists available at ScienceDirect

Stem Cell Research

journal homepage: www.elsevier.com/locate/scr

The evolution of Great Apes has shaped the functional enhancers' landscape in human embryonic stem cells

Gennadi Glinsky[a,*], Tahsin Stefan Barakat[b]

[a] Institute of Engineering in Medicine, University of California San Diego, 9500 Gilman Dr. MC 0435, La Jolla, CA 92093-0435, USA
[b] Department of Clinical Genetics, Erasmus MC, University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands

ABSTRACT

High-throughput functional assays of enhancer activity have recently enabled the genome-scale definition of molecular, structural, and biochemical features of these genomic regulatory regions. To infer the evolutionary origin of DNA sequences operating as functional enhancers in human embryonic stem cells (hESC), we examined the patterns of evolutionary conservation and divergence in the genome-wide functional enhancers' landscape of hESC. We show that a prominent majority (up to 94%) of DNA sequences identified in hESC as functional enhancers are conserved in humans and our closest evolutionary relatives, Chimpanzee and Bonobo. More than 91% of functional enhancers that are highly conserved in both Chimpanzee and Bonobo, are conserved among other Great Apes and > 75% are conserved in the Rhesus genome. In striking contrast, < 5% of DNA sequences operating in hESC as functional enhancers are conserved in rodents. Conserved in primates enhancers' sequences are complemented by 1619 sequences of enhancers that are specific to humans. Enhancers that harbor human-specific sequences appear enriched among the invariant enhancer module maintaining activity in different pluripotent states and these regions are associated with pluripotency- and embryonic-lineage-related genes. However, functional enhancers make up only a minority of all conserved in primates or human-specific transcription factor binding sites. Our analyses revealed that sequences that are conserved during ~8 million years of primate evolution dominate the genomic landscape of functional enhancers in both primed and naïve hESC. Collectively, these observations revealed thousands of evolutionarily conserved sequences that function as a core regulatory network in human embryonic stem cells which has recently undergone further extension after divergence of modern humans from our closest relatives, Chimpanzee and Bonobo.

1. Introduction

Studies into changes in the DNA during evolution have helped to shape our understanding of mechanisms of phenotypic changes. However, despite recent progress and a comprehensive catalogue of changes within protein-coding genes, the genetic basis of many divergent features remains elusive (Chimpanzee Sequencing and Analysis Consortium, 2005; Fu et al., 2014; Green et al., 2010; Meyer et al., 2012; Prufer et al., 2012; Prufer et al., 2014), thus supporting the hypothesis that the evolution of regulatory loci in the genome have contributed to the emergence of unique human phenotypes (King and Wilson, 1975).

An extensive search for human-specific genomic regulatory loci has resulted in identification of ~20,000 candidates, the vast majority of which is located within non-coding regions of the human genome (Capra et al., 2013; Glinsky, 2015a, 2016a,b,c, 2017, 2018a,b; Konopka et al., 2012; Marnetto et al., 2014; McLean et al., 2011; Shulha et al., 2012). Most recently, the analyses of individual genomes of Great Apes using high-resolution sequencing technologies (Kronenberg et al., 2018) and applications of novel bioinformatics approaches to genome-wide expression profiling of human prefrontal cortex (Guffanti et al., 2018) markedly expanded the compendium of candidate human-specific regulatory sequences, which currently comprises nearly sixty thousand genomic loci aligned to the most recent release of the human reference genome (Glinsky, 2018b).

Transposable elements (TEs) represent a major evolutionary force contributing to the creation of species-specific regulatory networks in primate genomes by providing highly efficient alternative promoters, novel transcription factor-binding sites, potent enhancers, as well as small- and long-non-coding RNAs ((Bourque et al., 2018; del Rosario et al., 2014; Jacques et al., 2013; Kelley and Rinn, 2012); and references therein). In human preimplantation embryogenesis and human embryonic stem cells (hESC), activation of regulatory loci derived from human endogenous retroviral sequences appears to play a critically important role (Durruthy-Durruthy et al., 2016; Fort et al., 2014; Glinsky, 2015a,b, 2016a,b,c, 2017, 2018a, Glinsky et al., 2018; Goke et al., 2015; Grow et al., 2015; Koyanagi-Aoi et al., 2013; Kunarso et al., 2010; Lu et al., 2014; Ohnuki et al., 2014; Santoni et al., 2012; Smith et al., 2014; Wang et al., 2014; Xue et al., 2013; Yan et al., 2013). We showed that sequences derived from TEs contribute to creation of thousands of these human-specific genomic regulatory loci (Glinsky, 2015a). Most loci are also bound by the transcription factors (TF)

* Corresponding author.
E-mail address: gglinskii@ucsd.edu (G. Glinsky).

https://doi.org/10.1016/j.scr.2019.101456
Received 30 June 2018; Received in revised form 23 April 2019; Accepted 30 April 2019
Available online 03 May 2019

NANOG, OCT4 (POU5F1), and CTCF proteins suggesting a regulatory role during early-stage embryogenesis and relevance to developmental disorders and cancer (Glinsky, 2015a,b, 2016a,b).

However, in previous analyses, the definition of regulatory regions, such as enhancers, was based only on sequence conservation, chromatin features such as histone modifications, chromatin accessibility, and TF binding (Ernst et al., 2011, 2016; Hoffman et al., 2013; Visel et al., 2007), and associated biochemical events (e.g., bi-directional transcription (Andersson et al., 2014; Melgar et al., 2011; Wu et al., 2014)), but lacked direct functional evidence and only select candidates were successfully validated (Andersson et al., 2014; Wu et al., 2014). In consequence, the functional significance of primate- and human-specific regulatory sequences remains unclear.

Recently, massively parallel reporter assays have been developed to address this bottleneck, providing a direct genome-wide functional readout of enhancer activity (Arnold et al., 2013; Barakat et al., 2018; Ernst et al., 2016). In our implementation, the discovery of active enhancers was made possible by combining chromatin immunoprecipitation with a reporter assay. Application of this technique, termed ChIP-STARR-seq, to human embryonic stem cells (hESCs) identified thousands of functional enhancers (Barakat et al., 2018) (resource available at http://hesc-enhancers.computational-epigenetics.org/). To maintain the pluripotent state, enhancers in hESC are bound by pluripotency TFs and associate with enhancer-specific histone marks. However, only a fraction of loci marked by NANOG, OCT4, H3K27ac and H3K4me1 function as enhancers in hESC. These observations highlight the importance of critically examining the association of enhancers with DNA segments identified as primate- and human-specific TFBS.

In this contribution, we show that conserved in primates' sequences harbor the majority of functional enhancers in both primed and naïve hESC (83.8% and 84.2%, respectively). Additionally, we identify a subset of 1619 enhancer sequences that is human-specific and which are associated with human-specific binding sites of the ESC-linked TFs NANOG and OCT4. We also note that 5697 enhancers that are active in different pluripotent states (*Invariant* module) are particularly enriched in human-specific sequences and that human-specific and conserved in primates enhancers in this module are associated with different biological roles. Our detailed exploration of the evolutionary framework of functional enhancers in hESC indicate that the transcriptional control of pluripotency underwent major changes during the evolution of primates and that this network continues to evolve after the divergence of modern humans from other Great Apes.

2. Results and discussion

2.1. Evolutionary origins of active enhancers in hESC

We focused our analysis on DNA sequences of active enhancers comprising of 32,353 and 36,417 functional regulatory elements in primed and naïve hESCs, respectively (Barakat et al., 2018) (Fig. 1a, Table 1). Within these functional enhancers, two modules of enhancers were distinguished: those overlapping ESC enhancers previously defined by signatures of TFBS and histone modifications (*Core* module) and those lacking these signatures (*Extended* module) (Barakat et al., 2018). We sought to further examine the potential functional significance of these enhancer modules by evaluating them in the context of evolutionary conservation. To this end, we defined enhancer sequences with at least 95% sequence identity during direct and reciprocal conversions between human, chimpanzee, and bonobo genomes, as conserved in primates (Fig. 1b). This definition indicates high sequence conservation during ~8 million years of evolution. Conserved in primates sequences constitute the majority of functional enhancer sequences in both pluripotent states examined (primed = 83.8%, naïve = 84.2%) with a notable difference ($p = 2.64 * 10^{-203}$; two-tailed Fisher's exact test) between the *Core* (93.8%) and *Extended* module (80.5%). On the other hand, 6.2% and 19.5% of active enhancers' sequences of the *Core* and the *Extended* modules, respectively, could be assigned to the divergent in humans' category (Table 1) comprising regulatory DNA sequences manifesting > 5% divergence after human/chimpanzee split.

In contrast, human DNA sequences that are highly conserved in rodents (as defined by direct and reciprocal conversions between human and mouse genomes at 95% sequence identity) represent only a small fraction (~4%) of hESC enhancers, regardless of the module assignment (Table 1). Thus, conserved sequences shared with rodents represent only a small fraction of hESC functional enhancers. Next, we identified the sequences of functional enhancers that could be defined as primate-specific because they are not present in the mouse genome using 10% cut-off of the sequence identity threshold (see Methods). Using this approach, we found that 28.3% of the *Core* and 39.7% of the *Extended* module enhancers cannot be mapped to the mouse genome using 10% level of sequence identity threshold (Fig. 1c). These findings indicate that a significantly higher proportion of *Extended* module enhancers in hESC originates from primate-specific sequences, e.g., sequences that are absent in the mouse genome (*Extended* module compared to *Core* module; $p = 1.89 * 10^{-77}$; two-tailed Fisher's exact test). Sizable fractions of functional enhancer sequences assigned to either Core module (~67%) or Extended module (~56%) manifest sequence conservation levels of > 10% and < 95% in the mouse genome. We reasoned that the potential functional significance of these intermediate patterns of sequence conservations is uncertain and evolutionary histories of functional enhancers' sequences manifesting these intermediate sequence conservation levels cannot be ascertained using the comparisons with rodent genomes. Because traces of their origin still can be found in the mouse genome, parental sequences may have been present in the genome of common ancestors and mutated to evolve into sequences operating as active enhancers in hESC.

In summary, our analyses point to a high sequence conservation of functional enhancer sequences over ~8 million years of primate evolution. Interestingly, despite overall similarity between primates, a subset of enhancers lacking strong associated chromatin marks (*Extended* module) manifests stronger changes compared to the *Core* module, indicating that these regulatory elements may be a relatively recent evolutionary addition to the functional enhancer network of hESCs.

2.2. Identification of human-specific sequences operating as active enhancers in hESC

We defined enhancer sequences not mapping to chimpanzee and bonobo as human-specific. This analysis identified a total of 1619 human-specific enhancers (Figs. 1b, 2a). Of these, 1034 were active in primed hESC (3.2% of total active enhancers) compared to 963 (2.7% of total active enhancers) in naive hESC. A substantial number ($n = 378$; retention rate = 39.3%) of human-specific enhancers show activity both in primed and naïve hESC (Table 2), which is a significantly higher proportion than that of all functional enhancers (retention rate = 18%; $p = 5.68 * 10^{-22}$; two-tailed Fisher's exact test).

As previously observed for primate-specific enhancers, differential abundance between *Core* and *Extended* module enhancers were also found for human-specific enhancers, with 0.5% and 4.1% of enhancers mapped to human-specific sequences, respectively ($p = 5.029 * 10^{-80}$; two-tailed Fisher's exact test; Fig. 2b). Similarly, sequences of active enhancers not conserved in non-human primates were more frequent in the *Extended* module ("divergent in humans sequences" = all sequences that are not conserved in primates; *Core* = 6.2%, *Extended* = 19.5%; $p = 2.64 * 10^{-203}$).

We conclude that the large set of conserved in primates enhancers in hESCs is complemented by 1619 human-specific enhancers, suggesting that regulatory networks governed by hESC enhancers underwent a major extension during primate evolution and further evolved after the

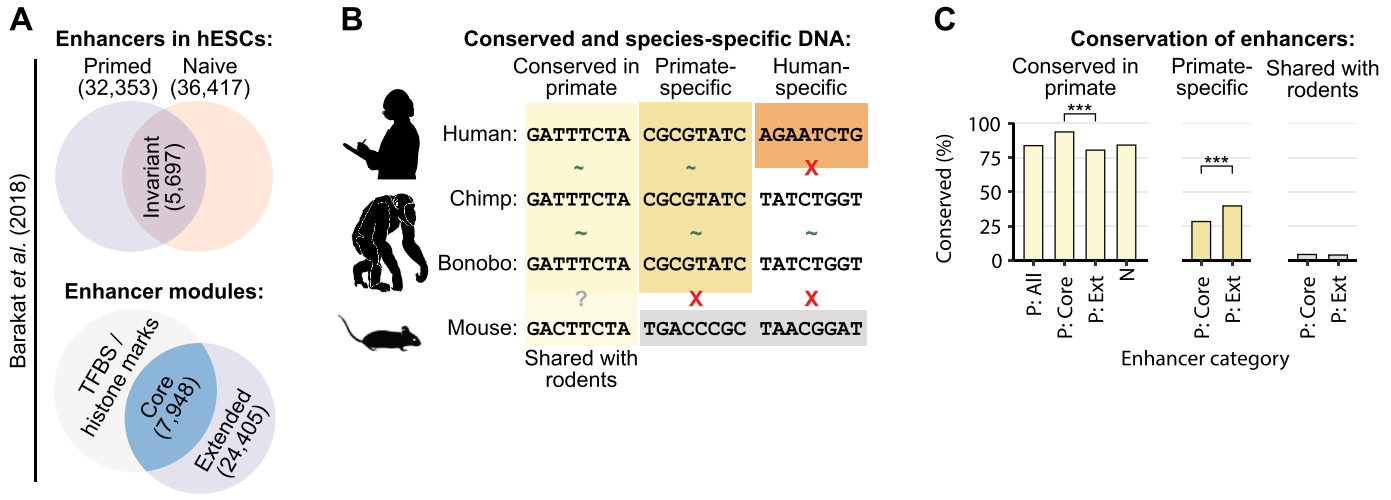

Fig. 1. Functional enhancers in human embryonic stem cells are dominated by conserved sequences.
(A) Schematic overview of functional enhancers identified in (Barakat et al., 2018) (Barakat et al., 2018).
(B) Schematic overview of conservation classes used in this study.
(C) Bar graph showing the percentage of enhancer sequences overlapping with different conservation classes. P: All: functional enhancers identified in primed hESCs. P: Core: functional enhancers identified in hESCs overlapping with the core enhancer module. P: Ext: functional enhancers identified in hESCs overlapping with the extended enhancer module. N: functional enhancers identified in naïve hESCs. TFBS: transcription factor binding sites.

divergence of humans.

2.3. Characterization of an invariant module of enhancers in primed and naive hESC

We next focused our analysis on the set of 5697 enhancers that retain activity in both primed and naïve hESC (*Invariant* module; Figs. 1a, 3). Consistent with our original results (Barakat et al., 2018), functional enrichment analyses with GREAT (McLean et al., 2010) confirmed that these enhancers were enriched near genes involved in the negative regulation of cell differentiation and developmental processes, as well as positive regulators of stem cell maintenance (Supplemental Table S1). Interestingly, we also found an enrichment of genes implicated in the post-implantation stages of embryonic development and early neurodevelopmental stages.

Conserved in primates sequences constitute 80.5% of the *Invariant* module, while another relatively high proportion (6.6%) are human-specific (a total of 378 regulatory sequences; Tables 3–5). The significant enrichment of human-specific enhancers in the *Invariant*

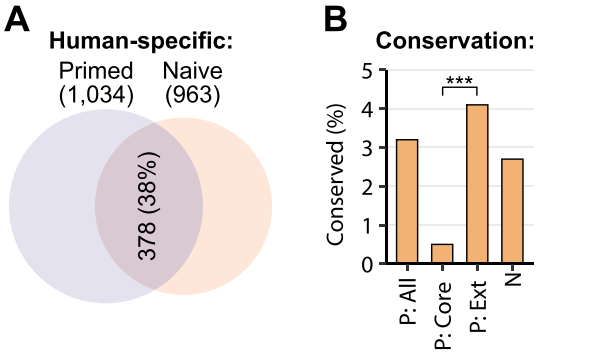

Fig. 2. Human-specific enhancer evolution in embryonic stem cells.
A) Venn diagram showing the overlap of human-specific active enhancer sequences in primed and naïve hESCs.
B) Bar diagram showing the conservation of human-specific sequences across different enhancer groups, as indicated in Fig. 1C.

Table 1
Evolutionary histories of the functional hESC enhancers assigned to distinct modules.

Classification category	Core module	Extended module	*P* value*	Primed hESC	Naïve hESC	*P* value*
hg19	7948	24,405		32,353	36,417	
hg38 (95% sequence identity conversion)	7945	24,318	0.862	32,263	36,333	0.966
hg38 conversion, %	99.96	99.64		99.72	99.77	
Primate-specific, n	2246	9652		11,898	12,830	
Primate-specific, %	28.27	39.69	6.11E-77	36.88	35.31	0.00002
Conserved in rodents, n	330	916		1246	1504	
Conserved in rodents, %	4.15	3.77	0.123	3.86	4.14	0.067
Conserved in primates, n	7455	19,572		27,027	30,601	
Conserved in primates, %	93.83	80.48	2.64E-203	83.77	84.22	0.108
Divergent in humans, n	490	4746		5236	5732	
Divergent in humans, %	6.17	19.52	2.64E-203	16.23	15.78	0.108
Human-specific, n	36	998		1034	963	
Human-specific, %	0.45	4.10	5.029E-80	3.20	2.65	0.0000169

*, *p* values were estimated using the 2-tailed Fisher's exact test; primate-specific, human-specific, conserved in rodents, and conserved in primates sequences were identified as described in the methods; primate-specific loci were defined when enhancer sequences do not intersect any chains in the mouse genome (mm10) with the sequence identity threshold of 10%; human-specific loci were defined when enhancer sequences do not intersect any chains in the genomes of both Chimpanzee and Bonobo with the sequence identity threshold of 10%; Numbers of enhancers in the divergent in humans category were calculated by subtracting the number of highly-conserved in primates enhancers from the total number of enhancers in the corresponding classification category.

Table 2
Increased numbers of human-specific naïve hESC functional enhancers share common genomic coordinates* with the extended module & core module functional enhancers of primed hESC.

Classification category	Naïve hESC (observed)	Naïve hESC (expected)**	Enrichment	*P* value***
Number of human-specific functional enhancers	963	NA		
Percent	100.00	NA		
ESC core module Primed hESC (n = 36)	20	6	3.33	0.0089
Percent	2.08	0.62		
Extended module Primed hESC (n = 998)	358	180	1.99	1.40E-19
Percent	37.18	18.69		
Extended & core modules Primed hESC (n = 1034)	378	186	2.03	5.68E-22
Percent	39.25	19.31		

*, Common genomic coordinates were defined as the identical chromosomal locations of the individual functional enhancers and the numbers of enhancers having the same chromosomal locations were computed for each functional enhancers' category; **, Expected number of enhancers in naïve hESC was estimated based on the overall fraction of 18% maintained in High-High functional enhancers' category in primed and naïve hESC; ***, p values were estimated using the 2-tail Fisher's exact test.

module was confirmed also in independent comparisons to the *Core* and *Extended* modules, and to active enhancers in naïve hESC (Tables 4–5). Functional enrichment analysis using Enrichr (Chen et al., 2013) showed that conserved in primates sequences where associated with many processes linked to hESC biology (such as transcriptional regulation of human pluripotent stem cells), whereas human-specific sequences also showed enrichment for other processes related to human reproduction and fertilization (Fig. 3B). Notably, conserved in primates components of the *Invariant* module are linked to genes down-regulated upon depletion of pluripotency factors (e.g., OCT4, SOX2, NANOG), whereas genes linked to the human-specific component are most affected by the depletion of global epigenetic regulators such as *TET1/2/3*. Genes linked to human-specific functional enhancers impact EPHA-mediated growth cone collapse and apoptosis, whereas conserved in primates component genes affect EPHB-mediated forward signaling and Sema4D induced cell migration and growth cone collapse (Fig. 3B).

Human-specific functional enhancer sequences are components of the more broadly-defined evolutionary category designated divergent in human sequences (Materials & Methods), which explain why they manifest similar patterns of enrichment (Tables 3–5). Notably, the enrichment of divergent in human sequences was not observed when the human-specific sequences were subtracted (Tables 4–5) indicating that the observed enrichment effects are linked to human-specific sequences operating in hESC as functional enhancers.

Overall, our analyses indicate that enhancers active in different pluripotent states (the *Invariant* module) underwent substantial changes after the divergence of humans and Great Apes.

2.4. Association of human-specific TF binding sites and active enhancers

In earlier work we defined human-specific (hsTFBS) and primate-specific (psTFBS) binding sites for pluripotency master TFs NANOG and

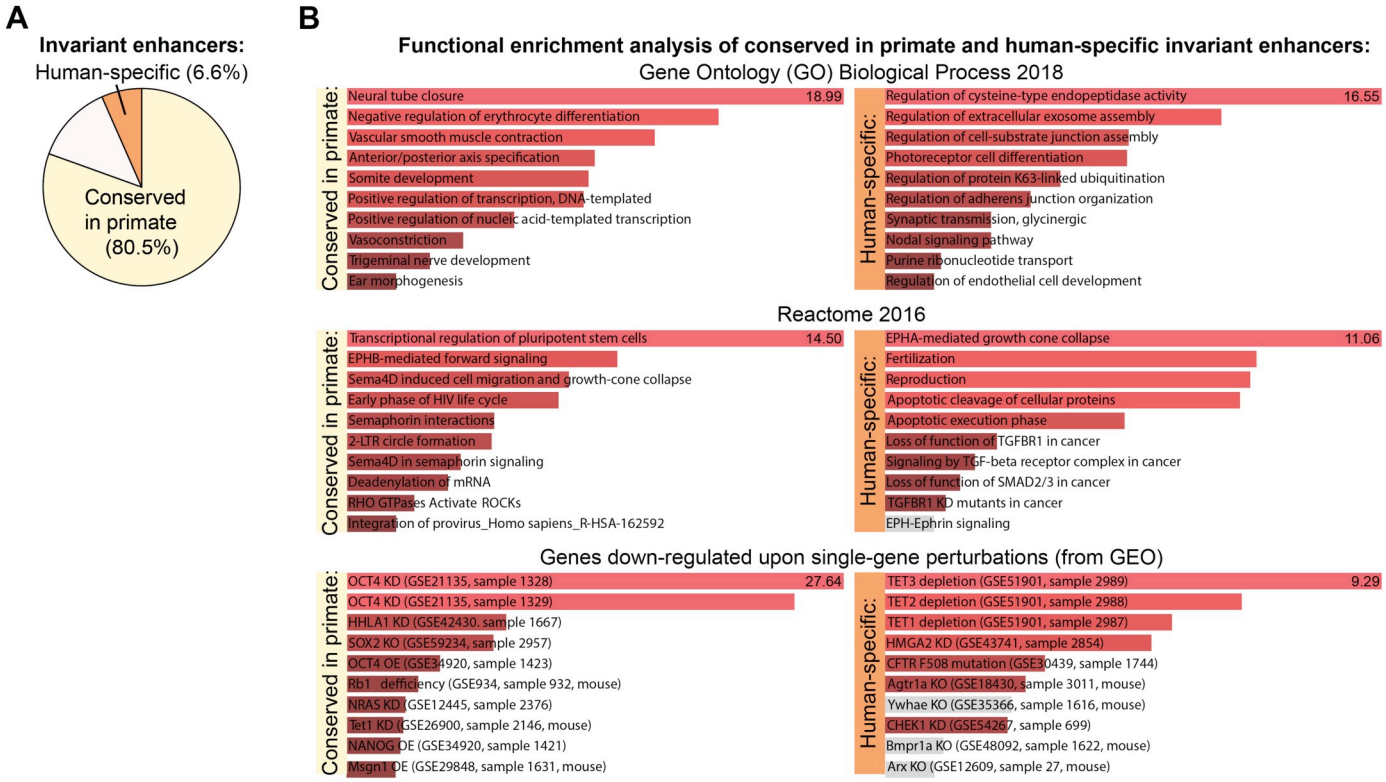

Fig. 3. Enhancers displaying activity in primed and naïve hESCs are composed of conserved in primates and human-specific sequences.
(A) Pie chart showing the fraction of invariant enhancers that are conserved in primates or human-specific sequences.
(B) Functional enrichment analysis using Enrichr (Chen et al., 2013) of conserved in primates and human-specific invariant enhancers.

Table 3
Invariant (master) module of hESC functional enhancers comprises 5697 enhancers maintaining common genomic coordinates* during transition from primed to naïve pluripotency state.

Classification category	Number of enhancers	Percent	Expected***	Enrichment	P value**
Conserved in primates active enhancers	4589	80.55	4772	0.96	8.36E-06
Extended module	2665	46.78	3456	0.77	5.35E-50
ESC core module	1924	33.77	1316	1.46	1.37E-36
Divergent in humans active enhancers	1108	19.45	925	1.20	8.36E-06
Extended module	971	17.04	838	1.16	0.0007
ESC core module	137	2.40	87	1.58	0.0009
Human-specific active enhancers	378	6.64	183	2.07	2.21E-17
Extended module	358	6.28	176	2.03	5.58E-16
ESC core module	20	0.35	6	3.15	0.0093
Invariant (Master) stemness module of active enhancers	5697	100.00	5697	1.00	NA
Extended module	3636	63.82	4294	0.85	5.20E-41
Core module	2061	36.18	1403	1.47	5.20E-41

*, Common genomic coordinates were defined as the identical chromosomal locations of the individual functional enhancers and the numbers of enhancers having the same chromosomal locations were computed for each functional enhancers' category; **, p values were estimated using the 2-tailed Fisher's exact test; ***, expected numbers of enhancers in each category were estimated based on the distribution of the corresponding enhancers' categories in the primed state hESC; numbers of enhancers in the divergent in humans category were calculated by subtracting the number of conserved in primates enhancers from the total number of enhancers in the corresponding classification category.

Table 4
Invariant stemness module of hESC functional enhancers is enriched for human-specific enhancers and evolutionary distinct categories of the core module functional enhancers.

Classification category	Invariant module	Percent	Primed hESC unique loci	Percent	Enrichment vs primed	P value*
Conserved in primates active enhancers	4589	80.55	22,438	84.46	0.95	1.11E-12
Extended module	2665	46.78	16,907	63.64	0.74	5.14E-121
Core module	1924	33.77	5531	20.82	1.62	6.60E-92
Divergent in humans** active enhancers	1108	19.45	4128	15.54	1.25	1.11E-12
Extended module	971	17.04	3775	14.21	1.20	7.32E-08
Core module	137	2.4	353	1.33	1.81	1.52E-08
Human-specific active enhancers	378	6.64	656	2.47	2.69	5.46E-49
Extended module	358	6.28	640	2.41	2.61	2.40E-44
Core module	20	0.35	16	0.06	5.81	3.31E-07
Divergent in humans without human-specific	730	12.81	3472	13.07	0.98	0.618
Extended module	613	10.76	3135	11.80	0.91	0.027
Core module	117	2.05	337	1.27	1.62	1.31E-05
Total number of enhancers	5697	100	26,566	100.00	1.00	
Extended module	3636	63.82	20,682	77.85	0.82	3.14E-103
Core module	2061	36.18	5884	22.15	1.63	3.14E-103

*, p values were estimated using the 2-tailed Fisher's exact test; **, numbers of enhancers in the divergent in humans' category were calculated by subtracting the number of conserved in primates' enhancers from the total number of enhancers in the corresponding classification category.

OCT4 by sequence conservation analyses (Glinsky, 2015a). However, it was previously not possible to assess the functional relevance of the evolutionary conservation of TF binding. For instance, psTFBS comprise only ~30%, and hsTFBS only 1%, of all TFBS in hESCs (Glinsky, 2015a; Kunarso et al., 2010), yet it is unknown what numbers of primate-specific and human-specific TFBS are associated with functional enhancers in hESC.

We consistently observed strong associations of human-specific functional enhancers with independently-defined hsTFBS (Tables 6, 7): About half of all human-specific functional enhancer sequences are located in close proximity (± 10 kb) to hsTFBS occupied by NANOG (primed = 47.5%, naïve = 44.3%) or OCT4 (primed = 56.4%, naïve = 49%) (Table 6). OCT4 and NANOG binding sites were enriched in our dataset of active enhancers by design of the reporter assay (which used DNA immunoprecipitated for OCT4 and NANOG as well as for the histone marks H3K27ac and H3K4me1 as inputs), but hsTFBS definitions were based on independent sets of laboratory experiments and computational analyses. It is thus reassuring to see this overlap. Importantly, the definition of human-specific sites is independent of both sets of experiments, using instead a lack of intersection with any chains in the reference genomes of both chimpanzee and bonobo. This definition indicates that these sequences are absent in genomes of our evolutionary closest relatives.

Detailed examination of different enhancer modules showed no strong differences, with 66.7% of human-specific sequences in the *Core* module and 61.5% of those in the *Extended* module mapping close to any hsTFBS (Table 7). Strikingly, a large proportion of human-specific enhancers (*Core* module = 50.0%; *Extended* module = 41.8%; naïve = 37.9%) are associated with overlapping binding sites for both NANOG and OCT4 (Fig. 4 and Table 7).

We further confirmed our results by stringent Genome-wide Proximity Placement Analysis (GPPA) of human-specific functional enhancers co-localizing with hsTFBS (co-localization is scored when the enhancer and TFBS occupy the exactly same genomic sequence), as done previously (Glinsky, 2015a, 2016a,b, 2018a). Briefly, this methods gauges the significance of overlaps between two types of genomic elements (here, active enhances and hsTFBS; see Methods and (Glinsky, 2015a, 2016a,b, 2018b). We consistently observed a significant co-localization of hsTFBS (NANOG, $p = 2.03 * 10^{-06}$; CTCF, $p = 1.46 * 10^{-08}$; OCT4, $p = 8.91 * 10^{-33}$) with functional enhancers that are active in both primed and naïve hESC (*Invariant* module; Table 8). Co-localizations of all hsTFBS (NANOG, $p = 5.01 * 10^{-19}$; CTCF, $p = 6.33 * 10^{-09}$; OCT4, $p = 5.04 * 10^{-20}$) with sequences of functional enhancers remain consistently significant when all instances of overlaps were scored (Table 8). In contrast to hsTFBS, psTFBS appear to manifest less consistent co-localization patterns with functional

Table 5
Invariant stemness module of hESC functional enhancers is enriched for human-specific enhancers compared to active enhancer sequences operating in either primed or naïve hESC.

Classification category	Invariant module	Percent	Primed state unique loci	Percent	Enrichment vs primed	P value	Naive state unique loci	Percent	Enrichment vs naive	P value	Enrichment vs primed	P value*
Conserved in primates active enhancers	4589	80.55	**22,438**	**84.46**	0.95	1.11E-12	26,012	84.91	0.95	6.12E-16	1.01	0.142
Divergent in humans active enhancers	1108	19.45	**4128**	**15.54**	1.25	1.11E-12	4624	15.09	1.29	6.12E-16	0.97	0.142
Human-specific active enhancers	378	6.64	**656**	**2.47**	2.69	5.46E-49	585	1.91	3.47	6.96E-72	0.77	5.32E-06
Divergent in humans without human-specific	730	12.81	**3472**	**13.07**	0.98	0.618	4039	13.18	0.97	0.691	1.01	0.455
Total number of enhancers	5697	100.00	**26,566**	**100.00**	1.00		30,636	100.00	1.00		1.00	

*, *p* values were estimated using the 2-tailed Fisher's exact test; **, numbers of enhancers in the divergent in humans' category were calculated by subtracting the number of conserved in primates' enhancers from the total number of enhancers in the corresponding classification category. Last two columns report the results of enrichment/depletion analyses of functional enhancers assigned to the Naïve state unique loci and Primed state unique loci categories; Functional enhancers' categories assigned to the Primed state unique loci are highlighted in bold.

enhancer sequences and in some instances significantly less co-localization events were observed than expected (Table 9).

Collectively, our analyses highlight a strong overlap of hsTFBS for master pluripotency regulators with human-specific functional enhancers in hESC. However, the majority of these hsTFBS (64–67%) was not active as enhancers, affirming that identification of hsTFBS alone is not sufficient for accurate enhancer prediction. Consistent with this conclusion, even larger fractions of primate-specific TFBS for OCT4 and NANOG (77–83%, respectively) did not to display enhancer activity (Table 9)

2.5. Not all conserved in primates sequences are functional enhancers

Our preceding analyses have clearly demonstrated that the majority of DNA sequences underlying active enhancers in hESCs has been highly conserved in humans, chimpanzee, and bonobo. We would like to underscore that high sequence conservation alone or in conjunction with enhancer-associated chromatin marks does not necessarily entail or reliably predict the enhancer activity. We had previously generated hESCs with by CRISPR-Cas9-mediated deletions of both active enhancer sequences ($n = 5$) and DNA sequences that had also been measured in our assay but were found to lack enhancer activity ($n = 6$). These experiments yielded altered target gene expression only for deletions of active enhancers (Barakat et al., 2018). All deleted control and enhancer regions were highly conserved in primate (100% of bases, 100% of span) and also had chromatin-based markers of putative enhancers (e.g., DNaseI clusters and TFBS for NANOG and OCT4) (Fig. 5 and Supplemental Figs. 1 and 2, and data not shown).

These experiments together with the additional analysis presented here established that evolutionary conservation alone or in combination with chromatin features are useful means for stratification and rationalization of enhancers, but must be combined with direct functional assays for reliable enhancer identification.

2.6. A majority of DNA segments operating as functional enhancers in hESC maintained sequence conservation during 30 million years of primate evolution

Our previous analyses identified DNA sequences that are highly conserved in humans, chimpanzee, and bonobo and manifest enhancers' activities in primed and naïve hESC (Table 1). This indicates that a prominent fraction of hESC functional enhancers originates from DNA sequences that retained at least 95% of sequence identity during ~8 million years of primates' evolution. It was of interest to extend the sequence conservation analyses of this highly conserved in primates' class of functional enhancers by including genomes of other non-human primates and rodents (Table 10). We included in this analysis the reference genomes of four non-human primates (Gorilla, Orangutan, Gibbon, and Rhesus) and two rodent species (mouse and rat). To enhance the stringency of the analytical pipeline, we also analyzed two additional independent reference genome databases of Chimpanzee (PanTro6) and Bonobo (PanPan1) that were not utilized in the previous analyses defining the evolutionary sequence conservation categories of functional enhancers (Table 10). Of importance is that the PanTro6 genome was generated without so called "humanized" genomic regions (Kronenberg et al., 2018), which gives a better estimation of conservation levels. The results of these analyses demonstrate that > 91% of functional enhancers' sequences that are highly conserved in genomes of humans, chimpanzee, and bonobo, are also highly conserved in the genomes of Gorilla and Orangutan. Furthermore, > 75% of DNA sequences representing this category of functional enhancers are highly conserved in the Rhesus genome as well. In striking contrast, only ~4% of these sequences are conserved in rodents' genomes (Table 10). Collectively, these observations highlight a significant contribution of the Great Apes' evolution to the genomic landscape of DNA sequences operating as functional enhancers in hESC and indicate that a majority of

Table 6
Genomic association* of human-specific hESC functional enhancers with human-specific NANOG and OCT4 (POU5F1) binding sites.

Human-specific NANOG binding sites (HSNBS)				
Classification category	Number of human-specific functional enhancers	Near HSNBS, n	Near HSNBS, %	Number of HSNBS**
Extended & core modules human-specific hESC enhancers	1034	491	47.5	227
Human-specific Naïve hESC active enhancers	963	427	44.3	200
Human-specific OCT4 binding sites (HSOBS)				
Classification category	Number of human-specific functional enhancers	Near HSOBS, n	Near HSOBS, %	Number of HSOBS**
Extended & core modules human-specific hESC enhancers	1034	583	56.4	1425
Human-specific Naïve hESC active enhancers	963	472	49.0	1258

HSNBS, human-specific NANOG binding sites; HSOBS, human-specific OCT4 binding sites; *, genomic associations were defined based on co-localizations of the HSNBS and/or HSOBS and human-specific hESC functional enhancers which were identified within the 10 Kb windows upstream and downstream of each enhancers' genomic coordinates in the hg38 release of the human reference genome; **, HSNBS and HSOBS associated with functional enhancers were defined based on the location of their genomic coordinates within ± 10 Kb windows of the corresponding enhancer genomic coordinates.

Table 7
Genomic association* of human-specific hESC functional enhancers with human-specific TFBS occupied by both NANOG and OCT4.

Associated human-specific TFBS	Number of functional enhancers	Percent
36 human-specific ESC module functional enhancers	24	66.67
Human-specific NANOG-binding sites	22	61.11
Human-specific OCT4-binding sites	20	55.56
Human-specific NANOG & OCT4 binding sites	18	50.00
Human-specific NANOG-binding sites only	4	11.11
Human-specific OCT4-binding sites only	2	5.56
998 human-specific extended module functional enhancers	614	61.52
Human-specific NANOG-binding sites	469	46.99
Human-specific OCT4-binding sites	563	56.41
Human-specific NANOG & OCT4 binding sites	417	41.78
Human-specific NANOG-binding sites only	52	5.21
Human-specific OCT4-binding sites only	145	14.53
963 human-specific Naïve hESC functional enhancers	534	55.45
Human-specific NANOG-binding sites	427	44.34
Human-specific OCT4-binding sites	472	49.01
Human-specific NANOG & OCT4 binding sites	365	37.90
Human-specific NANOG-binding sites only	62	6.44
Human-specific OCT4-binding sites only	107	11.11

HSNBS, human-specific NANOG binding sites; HSOBS, human-specific OCT4 binding sites; *, genomic associations were defined based on co-localizations of the HSNBS and/or HSOBS and human-specific hESC functional enhancers which were identified within the 10 Kb windows upstream and downstream of each enhancers' genomic coordinates in the hg38 release of the human reference genome; HSNBS and HSOBS associated with functional enhancers were defined based on the location of their genomic coordinates within ± 10 Kb windows of the corresponding enhancer genomic coordinates.

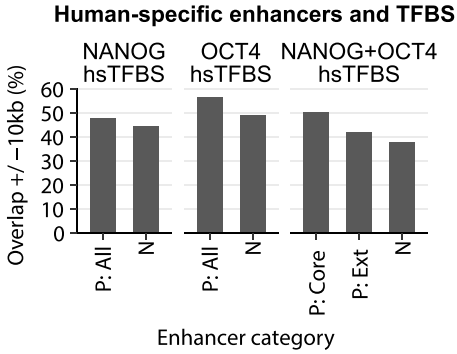

Fig. 4. Human-specific enhancers and transcription factor binding sites.
Bar graph showing the percentage of enhancers from different categories (as in Fig. 1C) that overlap with human-specific NANOG and OCT4 binding sites, or with binding sites co-occupied by both NANOG and OCT4. See text and Tables 6 and 7 for details.

functional enhancers maintained at least 95% of sequence conservation during ~30 million years of primates' evolution.

2.7. Distinct patterns of evolutionary histories of hESC functional enhancers and candidate hESC enhancers defined based on chromatin marks

To gain further insights into evolutionary origins of DNA segments operating in hESC as enhancers, we compared the profiles of evolutionary categories of hESC functional enhancers and 7006 sequences defined as putative hESC enhancers based on chromatin marks, including 684 super-enhancers (Supplemental Tables S2 and S3) (Hnisz et al., 2013). Because chromatin marks-defined putative enhancers were identified in primed hESC (Hnisz et al., 2013), the comparisons were made with primed hESC functional enhancers. These analyses demonstrate that chromatin marks-defined putative enhancers' sequences and DNA sequences functioning in hESC as active enhancers manifest significant differences of the quantitative balance of distinct evolutionary categories (Supplemental Table S2). The most apparent and consistent differences were the significantly increased fractions of primate-specific (p = 8.16E-75) and human-specific (p = 8.89E-67) sequences operating in primed hESC as functional enhancers compared to 7006 chromatin marks-defined putative enhancers (Supplemental Table S2, last column). This pattern of differences was also observed when comparisons were made of functional enhancers' sequences with DNA sequences of either the 6322 conventional enhancers or 684 super-enhancers separately (Supplemental Table S2). Notably, DNA segments defined as super-enhancers manifest a significantly lower fraction of highly-conserved in primates' sequences compared to either conventional chromatin marks-defined putative enhancers (p = 5.42E-33; Supplemental Table S2) or hESC functional enhancers (p = 3.39E-27; 2-tailed Fisher's exact test). Analysis of genomic coordinates of super-enhancers and functional enhancers revealed that 476 of 684 (69.6%) of DNA regions defined as super-enhancers contain at least one DNA segment operating as active enhancers identified in primed hESC (Supplemental Table S3). Taking into account functional enhancer sequences operating in both primed and naïve hESC, a total of 579 of 684 (84.7%) of super-enhancers harbor active enhancer sequences (Supplemental Table S3). A smaller fraction of conventional chromatin marks-defined enhancers (1680 candidate enhancers; 26.6%) overlap with functional enhancer sequences identified in primed hESC. Next, we confirmed differences in evolutionary histories of enhancers' sequences defined by different methodological approaches by performing the sequence conservation analyses only on overlapping enhancers' sequences, thus restricting the analyses to 2156 chromatin marks-defined enhancers that harbor 3463 functional enhancers identified in primed hESC (Supplemental Table S4). Collectively, the results of these analyses indicate that DNA segments defined as functional enhancers in hESC represent evolutionary-conserved sequences operating as active

Table 8
Co-localization of human-specific transcription factor binding sites with active enhancers (RPP > 138) in naïve and primed hESC.

Classification category	Number of scored loci	Percent	Expected	Observed/expected ratio**	P value*
Human-specific NANOG-binding sites (n = 826)	803	97.22			
Naïve hESC Active Enhancers	97	35.53			
Naïve & Primed hESC Active Enhancers	99	36.26	49	2.02	2.03E-06
Primed hESC Active Enhancers	77	28.21			
NANOG all Active Enhancers	273	100.00	120	2.28	5.01E-19
Human-specific OCT4-binding sites (n = 2386)	2277	95.43			
Naïve hESC Active Enhancers	198	23.13			
Naïve & Primed hESC Active Enhancers	381	44.51	154	2.47	8.91E-33
Primed hESC Active Enhancers	277	32.36			
OCT4 all Active Enhancers	856	100.00	569	1.50	5.04E-20
Human-specific CTCF-binding sites (n = 591)	564	95.43			
Naïve hESC Active Enhancers	41	20.40			
Naïve & Primed hESC Active Enhancers	89	44.28	36	2.47	1.46E-08
Primed hESC Active Enhancers	71	35.32			
CTCF all Active Enhancers	201	100.00	113	1.78	6.33E-09

*, p values were estimated using the 2-tailed Fisher's exact test; **, expected numbers of co-localized loci of transcription factor binding sites and active enhancers were estimated at 15% for NANOG; 25% for OCT4; and 20% for CTCF; expected numbers of Naïve & Primed active enhancers were estimated at 18% (Barakat et al., 2018).

Table 9
Co-localization of primate-specific transcription factor binding sites with active enhancers (RPP > 138) in naïve and primed hESC.

Classification category	Number of scored loci	Percent	Expected	Observed/expected ratio**	P value*
Primate-specific NANOG-binding sites (n = 28,843)	28,843	100.00			
Naïve hESC Active Enhancers	2228	45.25			
Naïve & Primed hESC Active Enhancers	885	17.97	886	1.00	1
Primed hESC Active Enhancers	1811	36.78			
NANOG all Active Enhancers	4924	100.00	4326	1.14	1.23E-11
Primate-specific OCT4-binding sites (n = 13,877)	13,877	100.00			
Naïve hESC Active Enhancers	1159	35.56			
Naïve & Primed hESC Active Enhancers	894	27.43	587	1.52	1.18E-19
Primed hESC Active Enhancers	1206	37.01			
OCT4 all Active Enhancers	3259	100.00	3469	0.94	3.42E-03
Primate-specific CTCF-binding sites (n = 28,626)	28,626	100.00			
Naïve hESC Active Enhancers	1031	45.64			
Naïve & Primed hESC Active Enhancers	187	8.28	407	0.46	2.02E-22
Primed hESC Active Enhancers	1041	46.08			
CTCF all Active Enhancers	2259	100.00	5725	0.39	0

*, p values were estimated using the 2-tailed Fisher's exact test; **, expected numbers of co-localized loci of transcription factor binding sites and active enhancers were estimated at 15% for NANOG; 25% for OCT4; and 20% for CTCF; expected numbers of Naïve & Primed active enhancers were estimated at 18% (Barakat et al., 2018).

enhancers within a majority of continuous linear genomic regions defined as super-enhancers, while only a small fraction of conventional chromatin marks-defined enhancers overlap functional enhancer sequences. Overall, active enhancers and putative functional genomic elements defined based on different chromatin features manifest similar conservation profiles with the highly conserved in primates sequences being the prevalent evolutionary category, while conserved in rodent sequences representing only a small minority (data not shown). These observations underscore the notion that the sequence conservation along or in combination with different chromatin marks is not sufficient to identify or reliably predict genomic sequences operating as active enhancers in hESC, in line with our previous findings (Barakat et al., 2018).

2.8. Implications of analyses of sequence conservation patterns of hESC functional enhancers

We carried out an analyses of evolutionary conservation patterns of different modules of active enhancers to infer their evolutionary history and compared them to other types of functional genomic elements (Tables 1–10; Supplemental Tables 2–4; and data not shown). Results of these analyses support the conclusion that genomic sequences operating in hESC as functional enhancers represent a collection of sequences with apparently different evolutionary histories, including a prominent majority of highly-conserved in primates sequences supplemented with a sizable number of human-specific DNA segments. Notably, DNA sequences conserved in rodents represent only a small fraction of genomic segments defined as active enhancers in hESC, strongly arguing that regulatory networks governing the stemness phenotype in hESC are vastly different from mouse ESC and predominantly comprise of regulatory elements created during primates' evolution. Interestingly, genes associated with conserved in primates and human-specific sequences of functional enhancers appear to contribute to biologically distinct functions (Fig. 3; Supplemental Table S1; and data not shown), suggesting that they may operate synergistically during human development. Overall, the conservation profiles of active enhancers defined by comparisons of their sequences in reference genomes of humans, non-human primates, and rodents appear similar to other functional genomic regions considered highly-conserved in evolution such as exons and DHS regions (data not shown). More broad comparisons of conservation patterns of defined functional genomic elements such as active enhancers to the genome-wide interspecies conservation are confounded by several factors. Estimates of genome-wide mutation rates established that they vary among different regions of the human genome (Harpak et al., 2016). Genome-wide mutation rates are also affected by multiple factors: for example, they are different between males and females (Li et al., 2002) and known to correlate with father's age (Kong et al., 2012). Genome-wide comparisons of the interspecies

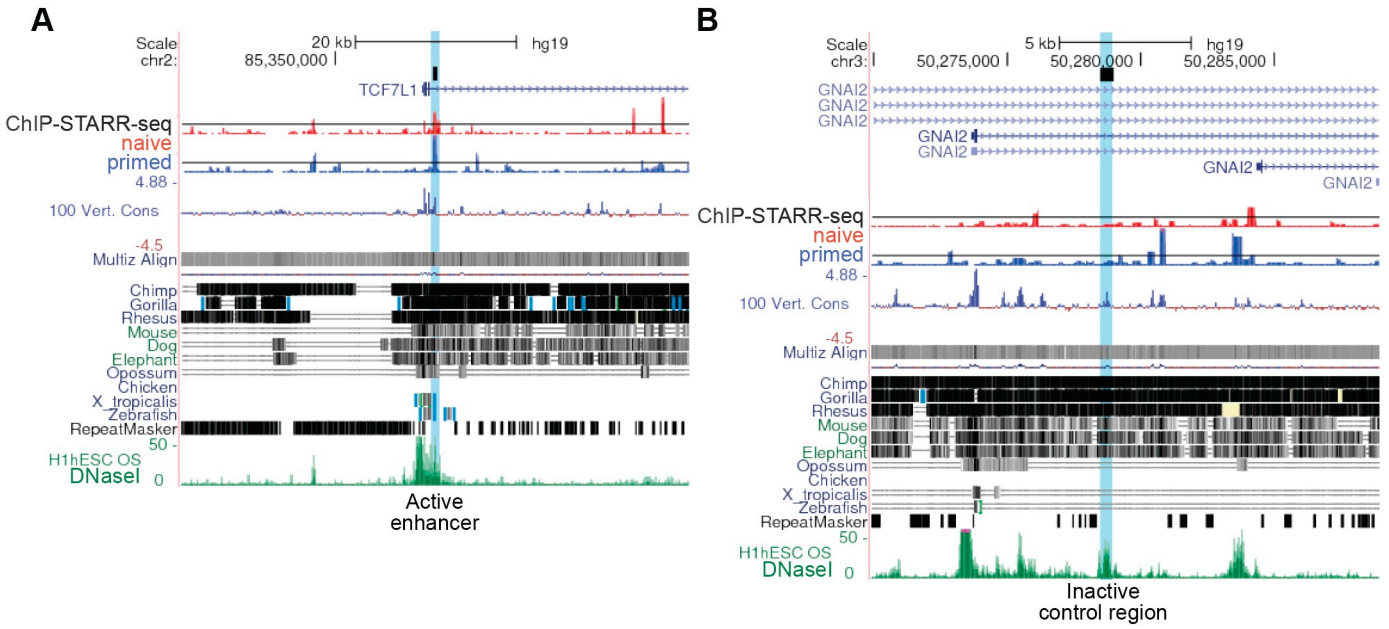

Fig. 5. Not all conserved sequences are functional enhancers in hESCs.
Exemplary Genome-browser shot of part of the TCF7L1 and GNAI2 loci. We previously generated cell lines with deletions of the indicated regions by CRISPR-Cas9 genome engineering. Both regions have a similar level of conservation and are marked by various enhancer features. In ChIP-STARR-seq analysis, only the TCF7L1 intronic sequences showed functional enhancer activity. Deletion of this conserved sequences in hESCs did affect gene expression of target genes, whereas the deletion of a similar conserved intronic region of GNAI2 that did not show activity in ChIP-STARR-seq did not affect target gene expression (Barakat et al., 2018).

conservations is also confounded by the known differences in the quality of the reference genome sequences for different species (with the human reference genome sequence quality being vastly superior: (Kronenberg et al., 2018)). Further complications arise from the approach called “humanization” of primates' reference genomes, which involves filing the sequence gaps with the human genome sequences of the orthologues regions. This problem was only recently resolved for some Great Apes reference genomes such as the PanTro6 database of the Chimpanzee genome (Kronenberg et al., 2018). Kronenberg et al. (2018) report that 83% of the ape genomes can be compared using a multiple sequence alignment and estimate that 36% of human autosomal DNA is subject to incomplete lineage sorting. To address these potential limitations, we extended our sequence conservation analyses to include the PanTro6 reference genome database released in 2018 (Table 10). Taking into account all these factors, in addition to the comparative analyses of different modules of active enhancers in both primed and naïve hESC, we performed the comparisons of the sequence conservation patterns of functional enhancers and chromatin marks-defined candidate enhancers (Supplemental Tables S2–S4). We observed consistent statistically significant differences between functional enhancers and chromatin marks-defined candidate enhancers reflecting a relatively higher fraction of evolutionary conserved sequences among genomic segments operating as active enhancers in hESC. Collectively, results of these analyses support the conclusion that the level of sequence conservation is not uniquely distinct feature of functional enhancers and combinations of sequence conservation analyses with any other chromatin-associated enhancers' features are not sufficient to distinguish and reliably identify functional enhancers.

3. Conclusions

Genomic screens for candidate enhancers typically utilize enhancer-associated molecular features such as histone marks, bidirectional transcription, and chromatin accessibility to define putative enhancers. These experimental approaches were recently extended by direct, genome-wide readouts of the functional activity of putative enhancers (Arnold et al., 2013; Barakat et al., 2018; Ernst et al., 2016). Here we have used the resource of enhancers in hESCs that we have recently made available (Barakat et al., 2018) to assess enhancer activity in the light of evolution.

Sequences that are highly conserved during ~8 million years of primate evolution were found to dominate the genomic landscape of functional enhancers in both primed and naïve hESC. Conserved in primates enhancers are complemented by a set of human-specific enhancers that are preferentially associated with human-specific TFBS of NANOG and OCT4. Conserved in primates enhancers are critical to regulate pluripotency related genes while human-specific enhancers appear to be also linked to other processes, such as human reproduction. These observations suggest that an essential genomic regulatory network for human embryonic stem cells was established during primate evolution to maintain a distinct from rodent primate stemness phenotype. Additionally, recent divergent evolution has resulted in an extended network to drive human-specific developmental traits in modern humans.

Interestingly, genes that are predictive of cancer survival across 17 human malignancies (Uhlen et al., 2017) are significantly enriched near DNA sequences operating as active hESC enhancers (data not shown), consistent with a role of stem cell regulatory networks in development of clinically lethal cancer phenotypes (Glinsky, 2015b, 2016a,c; Glinsky et al., 2005). This observation stresses the relevance of the examination of hESC functional enhancers not only for developmental and stem cell biology, but also to human disease. Collectively, our analyses illustrate that active enhancers operating in primed and naïve hESC represent complimentary sets of a tractable model for the interrogation of precise structure-activity-phenotype relations and a resource for dissecting the genetic elements governing evolutionarily stable and species-specific features of development, physiology and pathology.

4. Materials and methods

4.1. Definitions of active enhancers in hESCs

Data about enhancer activity in H9 human embryonic stem cells was obtained from Barakat et al. (Barakat et al., 2018). Briefly, this dataset

Table 10
Distinct conservation patterns among Great Apes and rodents of hESC functional enhancers operating on DNA sequences conserved in genomes of humans, chimpanzee, and bonobo.

Genome database	Species	Number of conserved enhancers	Percent
10.1. Conservation among Great Apes of 7455 core module sequences of primed hESC functional enhancers that are highly conserved in humans (hg38), chimpanzee (PanTro5), and bonobo (PanPan2).			
PanTro6	Chimpanzee	7424	99.58
PanPan1	Bonobo	7422	99.56
GorGor5	Gorilla	7300	97.92
PonAbe3	Orangutan	6966	93.44
NomLeu3	Gibbon	6207	83.26
RheMac8	Rhesus	5892	79.03
rn6	Rat	287	3.85
mm10	Mouse	328	4.40
10.2. Conservation among Great Apes of 19,572 extended module sequences of primed hESC functional enhancers that are highly conserved in humans (hg38), chimpanzee (PanTro5), and bonobo (PanPan2)			
PanTro6	Chimpanzee	19,489	99.58
PanPan1	Bonobo	19,480	99.53
GorGor5	Gorilla	18,929	96.71
PonAbe3	Orangutan	17,723	90.55
NomLeu3	Gibbon	14,944	76.35
RheMac8	Rhesus	14,711	75.16
Rn6	Rat	800	4.09
mm10	Mouse	809	4.13
10.3. Conservation among Great Apes of 30,601 sequences of naive hESC functional enhancers that are highly conserved in humans (hg38), chimpanzee (PanTro5), and bonobo (PanPan2)			
PanTro6	Chimpanzee	30,465	99.56
PanPan1	Bonobo	30,490	99.64
GorGor5	Gorilla	29,727	97.14
PonAbe3	Orangutan	27,987	91.46
NomLeu3	Gibbon	23,881	78.04
RheMac8	Rhesus	23,497	76.79
rn6	Rat	1356	4.43
mm10	Mouse	1375	4.49

Conservation patterns among non-human primates and rodents were evaluated for hESC functional enhancers operating on DNA sequences that are conserved in genomes of humans, chimpanzee, and bonobo (Table 1). Numbers of enhancers' sequences manifesting at least 95% sequence identity conservation during direct & reciprocal conversions from/to hg38 human reference genome database are reported for each species.

assessed the enhancer activity of 362,358 regions based on a massively parallel reporter assay using DNA immunoprecipitated for OCT4, NANOG, H3K27ac, or H3K4me1 as input. 32,353 and 36,417 regions were identified as active enhancers in primed and naïve hESCs, respectively. Active enhancers in primed hESCs were further distinguished into two modules, *Core* and *Extended*. *Core* module enhancers overlapped with at least one of 76,666 putative enhancers previously predicted based on TF binding or histone modification codes, while enhancers assigned to the *Extended* module lacked this overlap (Barakat et al., 2018).

4.2. Categories of DNA sequence conservation

Identification of conserved in primates, primate-specific, and human-specific sequences was performed as previously described (Glinsky, 2015a, 2016a,b, 2018a). In brief, all categories were defined by direct and reciprocal mapping using liftOver (Hinrichs et al., 2006). Specifically:

- *Conserved in primates*: DNA sequences that have at least 95% of bases remapped during conversion from/to human (*Homo sapiens*, hg38), chimp (*Pan troglodytes*, v5), and bonobo (*Pan paniscus*, v2).
- *Primate-specific*: Functional enhancers' sequences, including conserved in primates sequences, which failed to map to the mouse genome (mm10).
- *Human-specific*: DNA sequences that failed to map at least 10% of bases from human to both chimp and bonobo genomes.

The human divergent category reflects the more broad classification, which is the number of features defined by the subtraction of highly-conserved in primates' enhancers (having sequence identity of at least 95% in genomes of humans, chimpanzee, and bonobo) from the total number of enhancers in the corresponding functional enhancers' category. This category has been independently determined for the core module; extended module; primed enhancers; naïve enhancers; and chromatin marks-defined candidate enhancers. Human-specific enhancers represent more restricted evolutionary category, which is defined by the sequences that are absent in genomes of both Chimpanzee and Bonobo using the sequence identity threshold of 10%. Thus, human-specific enhancers are included in the more broad divergent in human category. To infer the putative evolutionary origins, each evolutionary classification was defined independently by running the corresponding analyses on all functional enhancers representing the specific category (the core module; extended module; primed enhancers; naïve enhancers; chromatin marks-defined enhancers). For example, human-rodent conversion identify sequences that are absent in the mouse genome based on the sequence identity threshold of 10%). Some highly-conserved in primates may be also conserved in rodents (in reality, only very few are shared with rodent). Non-rodent enhancers are best defined by the primate-specific classification (absent in the mouse genome based on the sequence identity threshold of 10%), which may include highly conserved in primates and human-specific sequences. However, highly conserved in primates and human-specific sequences represent non-overlapping evolutionary classifications. Similarly, highly conserved in primates and human divergent categories are non-overlapping by definition. Additional comparisons were performed using the same methodology and exactly as stated in the manuscript text.

4.3. Genome-wide proximity placement analysis

Genome-wide Proximity Placement Analysis (GPPA) of human-specific enhancers co-localizing with hsTFBS was carried out as described previously (Glinsky, 2015a, 2016a,b, 2018b). Briefly, we examined the significance of overlaps between active enhances and hsTFBS by first identifying all hsTFBS that overlap with any of the genomic regions tested in our ChIP-STARR-seq dataset. We then calculated the relative frequency of active enhancers overlapping with hsTFBS (Table 8). To assess the significance of the observed overlap of genomic coordinates, we compared the values recorded for hsTFBS with the expected frequency of active and non-active enhancers that overlap with all TFBS for NANOG (15%) and OCT4 (25%) as previously determined (Barakat et al., 2018). Our analyses demonstrate that > 95% of hsTFBS co-localized with sequences in the tested regions of the hESC genome (Table 8). GPPA allows the evaluation of both direct overlap and co-localization of defined genomic elements within DNA segments of specified continuous lengths.

4.4. Functional enrichment analysis

We used the GREAT web interface (version 3.0.0) (http://great.stanford.edu/public/html/) (McLean et al., 2010) for gene ontology analysis, using the following settings: basal plus extension, proximal 5 kb upstream and 1 kb downstream, plus distal up to 100 kb, including curated regulatory domains, and whole genome (hg19) as background. In addition, we used the Enrichr API (January 2018 version) (Chen et al., 2013) to test genes linked to enhancers of interest for significant enrichment in numerous functional categories. To comply with the web interface, we considered the 1000 genes closest to the tested peaks for enrichments. In all plots, we report the “combined score” calculated by

Enrichr, which is a product of the significance estimate and the magnitude of enrichment (combined score $c = log(p) * z$, where p is the Fisher's exact test p-value and z is the z-score deviation from the expected rank).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scr.2019.101456.

Acknowledgements

We thank Florian Halbritter (Vienna) for helpful discussion. We are indebted to the three anonymous reviewers for their constructive feedback. TSB is supported by the Netherlands Organisation for Scientific Research (ZonMW Veni, grant 91617021) and by an NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation.

Author contributions

GG conceived the study, and performed most of the conservation analyses. TSB performed functional enrichment analysis. GG and TSB interpreted results and wrote the paper.

Appendix A. Supplementary data

Supplemental Figs. 1 and 2 show the zoom-in views of CRISPR-Cas9-targeted loci and precisely displaying the enhancers-like chromatin features of deleted genomic regions.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al., 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., Stark, A., 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074–1077.

Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C., Chambers, I., 2018. Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell *Cell Stem Cell* 23, 276–288. 2018. *e8*. https://doi.org/10.1016/j.stem.2018.06.014 *(Epub 2018 Jul 19)*.

Bourque, G., et al., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199.

Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L., Pollard, K.S., 2013. Many human accelerated regions are developmental enhancers. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 368, 20130025.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128.

Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

del Rosario, R.C., Rayan, N.A., Prabhakar, S., 2014. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. Genome Res. 24, 1469–1484.

Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E.J., Davila, J., Mall, M., Wong, W.H., Wysocka, J., Au, K.F., Reijo Pera, R.A., 2016. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat. Genet. 48, 44–52.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. 34, 1180–1190.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., Noro, Y., Wong, C.H., de Hoon, M., Andersson, R., Sandelin, A., Suzuki, H., Wei, C.L., Koseki, H., FANTOM Consortium, Hasegawa, Y., Forrest, A.R., Carninci, P., 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nature Genet 46, 558–566.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prufer, K., de Filippo, C., et al., 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514, 445–449.

Glinsky, G.V., 2015a. Transposable elements and DNA methylation create in embryonic stem cells human-specific regulatory sequences associated with distal enhancers and noncoding RNAs. Genome Biol. Evol. 7, 1432–1454.

Glinsky, G.V., 2015b. Viruses, stemness, embryogenesis, and cancer: a miracle leap toward molecular definition of novel oncotargets for therapy-resistant malignant tumors? Oncoscience 2, 751–754.

Glinsky, G.V., 2016a. Activation of endogenous human stem cell-associated retroviruses (SCARs) and therapy-resistant phenotypes of malignant tumors. Cancer Lett. 376, 347–359.

Glinsky, G.V., 2016b. Mechanistically distinct pathways of divergent regulatory DNA creation contribute to evolution of human-specific genomic regulatory networks driving phenotypic divergence of Homo sapiens. Genome Biol. Evol. 8, 2774–2788.

Glinsky, G.V., 2016c. Single cell genomics reveals activation signatures of endogenous SCAR's networks in aneuploid human embryos and clinically intractable malignant tumors. Cancer Lett. 381, 176–193.

Glinsky, G.V., 2017. Human-specific features of pluripotency regulatory networks link NANOG with fetal and adult brain development. In: BioRxiv, https://doi.org/10.1101/022913. https://www.biorxiv.org/content/early/2017/06/19/022913.

Glinsky, G.V., 2018a. Contribution of transposable elements and distal enhancers to evolution of human-specific features of interphase chromatin architecture in embryonic stem cells. Chromosom. Res. 26, 61–84.

Glinsky, G.V., 2018b. Multi-species mosaicism of evolutionary origins of genomic loci harboring 59,732 human-specific regulatory sequences reflects a complex continuous speciation process of the human lineage. BioRxiv. https://doi.org/10.1101/432625. https://www.biorxiv.org/content/early/2018/10/03/432625.

Glinsky, G.V., Berezovska, O., Glinskii, A.B., 2005. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J. Clin. Invest. 115, 1503–1521.

Glinsky, G., Durruthy-Durruthy, J., Wossidlo, M., Grow, E.J., Weirather, J.L., Au, K.F., Wysocka, J., Sebastiano, V., 2018. Single cell expression analysis of primate-specific retroviruses-derived HPAT lincRNAs in viable human blastocysts identifies embryonic cells co-expressing genetic markers of multiple lineages. Heliyon 4, e00667. https://doi.org/10.1016/j.heliyon.2018.e00667. (eCollection 2018 Jun. PMID: 30003161).

Goke, J., Lu, X., Chan, Y.S., Ng, H.H., Ly, L.H., Sachs, F., 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.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722.

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., Pera, R.A., Wysocka, J., 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–225.

Guffanti, G., Bartlett, A., Klengel, T., Klengel, C., Hunter, R., Glinsky, G., Macciardi, F., 2018. Novel bioinformatics approach identifies transcriptional profiles of lineage-specific transposable elements at distinct loci in the human dorsolateral prefrontal cortex. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msy143.

Harpak, A., Bhaskar, A., Pritchard, J.K., 2016. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genet. 12, e1006489.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al., 2006. The UCSC genome browser database: update 2006. Nucleic Acids Res. 34, D590–D598.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., Young, R.A., 2013. Super-enhancers in the control of cell identity and disease. Cell 155, 934–947.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al., 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841.

Jacques, P.É., Jeyakani, J., Bourque, G., 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 9, e1003504.

Kelley, D., Rinn, J., 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 13, R107.

King, M.C., Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. Science 188, 107–116.

Kong, A., et al., 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471–475.

Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M.C., Gao, F., Chen, L., Wang, G.Z., Luo, R., Preuss, T.M., et al., 2012. Human-specific transcriptional networks in the brain. Neuron 75, 601–617.

Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K., Okita, K., Noma, H., Sawamura, Y., Teramoto, I., Narita, M., Sato, Y., Ichisaka, T., Amano, N., Watanabe, A., Morizane, A., Yamada, Y., Sato, T., Takahashi, J., Yamanaka, S., 2013. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc. Natl. Acad. Sci. U. S. A. 110, 20569–20574.

Kronenberg, Z.N., et al., 2018. High-resolution comparative analysis of great ape genomes. Science 360, eaar6343.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., Bourque, G., 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. 42, 631–634.

Li, W.-H., Yi, S., Makova, K., 2002. Male-driven evolution. Curr. Opin. Genet. Dev. 12, 650–656.

Lu, X., Sachs, F., Ramsay, L., Jacques, P.É., Göke, J., Bourque, G., Ng, H.H., 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat. Struct. Mol. Biol. 21, 423–425.

Marnetto, D., Molineris, I., Grassi, E., Provero, P., 2014. Genome-wide identification and characterization of fixed human-specific regulatory regions. Am. J. Hum. Genet. 95, 39–48.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M.,

Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501.
McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., Schaar, B.T., et al., 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471, 216–219.
Melgar, M.F., Collins, F.S., Sethupathy, P., 2011. Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biol. 12, R113.
Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C., et al., 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226.
Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., Nakamura, M., Tokunaga, Y., Nakamura, M., Watanabe, A., Yamanaka, S., Takahashi, K., 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc. Natl. Acad. Sci. U. S. A. 111, 12426–12431.
Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., et al., 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486, 527–531.
Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49.
Santoni, F.A., Guerra, J., Luban, J., 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology 9, 111.
Shulha, H.P., Crisci, J.L., Reshetov, D., Tushir, J.S., Cheung, I., Bharadwaj, R., Chou, H.J., Houston, I.B., Peter, C.J., Mitchell, A.C., et al., 2012. Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. PLoS Biol. 10, e1001427.
Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., Meissner, A., 2014. DNA methylation dynamics of the human preimplantation embryo. Nature 511, 611–615.
Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al., 2017. A pathology atlas of the human cancer transcriptome. Science 357.
Visel, A., Minovitsky, S., Dubchak, I., Pennacchio, L.A., 2007. VISTA enhancer browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 35, D88–D92.
Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V., Schumann, G.G., Chen, W., Lorincz, M.C., Ivics, Z., Hurst, L.D., Izsvák, Z., 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 516, 405–409.
Wu, H., Nord, A.S., Akiyama, J.A., Shoukry, M., Afzal, V., Rubin, E.M., Pennacchio, L.A., Visel, A., 2014. Tissue-specific RNA expression marks distant-acting developmental enhancers. PLoS Genet. 10, e1004610.
Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J.Y., Horvath, S., Fan, G., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597.
Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., Tang, F., 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131–1139.