Limits of sequence and functional conservation

Len A Pennacchio & Axel Visel

Sequence conservation of noncoding DNA across species can indicate functional conservation. However, a new study demonstrates notable differences between human and mouse stem cell regulatory networks, suggesting caution in generalizing from sequence to functional conservation.

Deciphering the gene regulatory architecture embedded in mammalian genomes is an essential prerequisite for understanding the role of regulatory sequences in human biology and disease. The identification of core sets of gene regulatory elements has been facilitated by cross-species sequence comparisons, but such sequence conservation–based approaches have limitations when exploring species-specific changes in gene regulation. On page 631 of this issue, Guillaume Bourque and colleagues[1] take an alternative approach, comparing the functional conservation, rather than the sequence conservation, of gene regulatory sites between the human and mouse genomes in embryonic stem (ES) cells. Remarkably, they find that the genomic locations of binding sites for two key regulatory proteins (OCT4 and NANOG) are poorly conserved across species, despite their functional importance in mammalian ES cell biology.

Functional divergence

Until now, most studies exploring the functional conservation of regulatory sequences across mammalian species have focused on experimental data sets obtained from only one species, followed by their *post hoc* comparative genomic analysis to infer degrees of DNA conservation across species[2–5]. These indirect studies have shown that some molecular marks associated with regulatory sequences tend to be found at sites whose sequence is highly conserved across species[4]. Other marks, however, tend to be found at sites with little or no sequence conservation[5], raising the question of to what extent such sites are functionally conserved across species. Kunarso *et al.*[1] now tackle this problem by obtaining genome-wide experimental data from both human and mouse by identical methodology.

To compare the genome-wide binding profiles of regulatory proteins between species, Kunarso *et al.*[1] performed ChIP-seq for three well-studied regulatory proteins (OCT4, NANOG and CTCF) from human and mouse ES cells. OCT4 (also known as POU5F1) and NANOG are transcription factors that play major roles in the maintenance of ES cell pluripotency, whereas the CTCF protein is associated with genomic insulator elements that prevent enhancer–promoter interactions. Unexpectedly, only ~5% of binding sites for the two transcription factors OCT4 and NANOG were found in orthologous positions in human and mouse ES cells, suggesting major differences in genome-wide binding profiles between species. Although subsets of these differences may be due to technical limitations of the approach, analysis of CTCF binding sites by identical methods revealed that, depending on statistical stringency, up to 50% of binding sites are functionally conserved between mouse and human. It is therefore reasonable to assume that the genome-wide binding profiles of OCT4 and NANOG in ES cells have substantially changed during the 75 million years of evolution that separate mice and humans from their last common ancestor.

Modes of regulatory rewiring

To elucidate the molecular mechanisms that led to the marked changes in the genome-wide binding profiles of OCT4 and NANOG in human compared to mouse ES cells, Kunarso *et al.*[1] examined the evolutionary origins of the sequences in which experimentally identified binding sites were located. Consistent with previous observations of regulatory sequences that arose through exaptation from transposable elements[6–9], between 10% and 30% of binding sites overlapped repeat elements (RABS, repeat-associated binding sites). Remarkably, many of these RABS were found in lineage-specific repeat elements that are absent in the comparison species, suggesting that large numbers of binding sites arose more recently in evolution and may have rewired the regulatory architecture in ES cells on a substantial scale.

To examine whether the changes in binding profiles functionally affect the transcriptional landscape of human and mouse ES cells, Kunarso *et al.*[1] quantified the impact and relative contributions of different modes of regulatory conservation and rewiring (**Fig. 1**). They obtained transcriptome-wide expression data from normal human ES cells, as well as from ES cells that had been depleted of OCT4 by RNA interference and compared these results to equivalent data from mouse ES cells. Overall, the genomic locations of OCT4 binding sites correlated with the locations of genes that were downregulated upon OCT4 depletion. However, among genes whose OCT4 dependence was conserved between human and mouse, most of the OCT4 binding sites identified were not directly conserved. Instead, the disappearance of a binding site in one species was compensated for by the emergence of a new binding site for the same transcription factor nearby. Kunarso *et al.*[1] further identified 50 cases in which human-specific OCT4 regulation could be directly linked to RABS—that is, cases of regulatory repeat-associated rewiring in human compared to mouse ES cells.

Function and annotation

Kunarso *et al.*[1] provide evidence that differences between the human and mouse ES cell transcriptomes are at least partially attributable to a divergence in genome-wide binding

Len A. Pennacchio and Axel Visel are at the Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA.
e-mail: lapennacchio@lbl.gov or avisel@lbl.gov

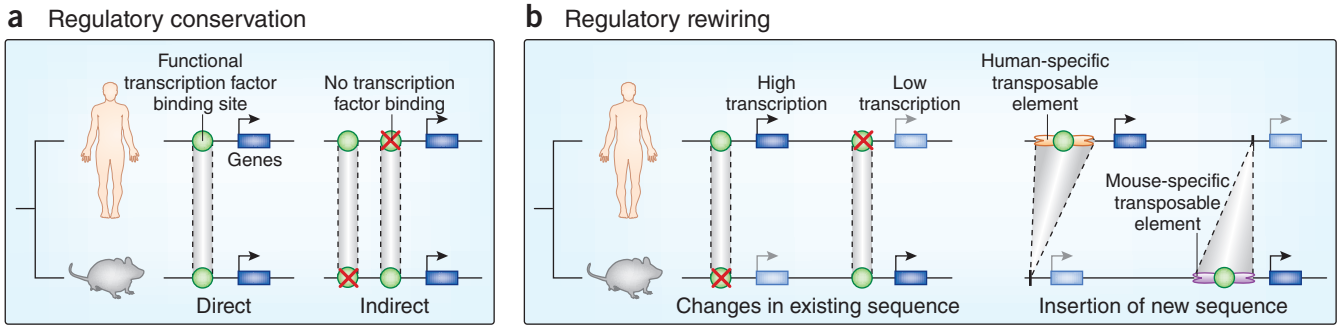

Figure 1 Conservation and rewiring of regulatory elements between human and mouse genomes. For a given transcription factor, binding site differences between species can have different effects on the transcriptional output of individual genes, resulting in either regulatory conservation or rewiring. (**a**) Transcriptional output is preserved (regulatory conservation) either through direct sequence conservation between species or through indirect conservation by creation of one binding site and destruction of another. (**b**) Transcriptional output is altered (regulatory rewiring) when binding sites are destroyed or created without compensatory nearby changes, or when new binding sites are created by insertion and exaptation of transposable elements.

profiles of major ES cell transcription factors. The study also provides insights into the unexpectedly large role that local binding site turnover, as well as RABS, play in the conservation and rewiring of mammalian regulatory networks. The functional relevance of the new human-specific OCT4 target genes identified by Kunarso *et al.*[1] remains to be determined, but they provide important leads for future studies.

Although sequence conservation has proven useful as a predictor of functional regulatory elements in the genome[2,10], the observations by Kunarso *et al.*[1] are a reminder that it is not justified to assume in turn that all functional regulatory elements show evidence of sequence constraint. It is noteworthy that whereas OCT4 binding and NANOG binding diverged between human and mouse ES cells, binding of CTCF was highly conserved. Thus, it is expected that other DNA-binding proteins and chromatin marks will fall into a spectrum from strong to weak conservation between these two species. The notion that some regulatory networks have substantially changed in evolution is also supported by recent independent observations of lineage-specific network rewiring in vertebrate preimplantation embryos and adult liver tissue[11,12]. The differences between species identified through these studies highlight the need to complement comparative genomic data with experimental approaches in order to obtain an accurate functional annotation of genomes.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Kunarso, G. *et al. Nat. Genet.* **42**, 631–634 (2010).
2. Cooper, G.M. & Brown, C.D. *Genome Res.* **18**, 201–205 (2008).
3. King, D.C. *et al. Genome Res.* **15**, 1051–1060 (2005).
4. Visel, A. *et al. Nature* **457**, 854–858 (2009).
5. ENCODE Project Consortium *et al. Nature* **447**, 799–816 (2007).
6. Bourque, G. *et al. Genome Res.* **18**, 1752–1762 (2008).
7. Wang, T. *et al. Proc. Natl. Acad. Sci. USA* **104**, 18613–18618 (2007).
8. Bejerano, G. *et al. Nature* **441**, 87–90 (2006).
9. Lowe, C.B., Bejerano, G. & Haussler, D. *Proc. Natl. Acad. Sci. USA* **104**, 8005–8010 (2007).
10. Visel, A., Rubin, E.M. & Pennacchio, L.A. *Nature* **461**, 199–205 (2009).
11. Xie, D. *et al. Genome Res.* **20**, 804–815 (2010).
12. Schmidt, D. *et al. Science* **328**, 1036–1040 (2010).

Hints of hidden heritability in GWAS

Greg Gibson

Although susceptibility loci identified through genome-wide association studies (GWAS) typically explain only a small proportion of the heritability, a classical quantitative genetic analysis now argues that considering together all common SNPs can explain a large proportion of the heritability of these complex traits. A related study provides recommendations for the sample sizes needed in future GWAS to identify additional susceptibility loci.

Genome-wide association studies (GWAS) have been highly successful in identifying genetic variants associated with hundreds of complex human traits and diseases, a feat that eluded two decades of linkage mapping. At the same time, the variants identified in GWAS generally only capture a few percent of the estimated heritability for these complex traits, leaving open the question of what may explain the remaining heritability. This may include contributions of rare variants, epistasis, epigenetics and genotype–environment interactions[1,2] but may also just imply that complex traits truly are affected by thousands of variants of small effect[3]. On page 565 of this issue, Peter Visscher and colleagues[4] report a joint estimate of the contribution of SNPs across all effect sizes and show that this can explain a large proportion of the heritability for height. On page 570 of this issue, Nilanjan Chatterjee and colleagues[5] examine recent GWAS for several complex traits and estimate the potential contribution of variants whose effect size is similar to that of discovered SNPs for diverse diseases. These two reports do not attempt to identify further genetic variants that could explain the remaining heritability, but they mount compelling arguments that a large proportion of it can be explained by common variants.

In GWAS, we test for association between the frequency of each of hundreds of thousands of common variants and a given phenotype, call

Greg Gibson is in the School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA.
email: greg.gibson@biology.gatech.edu

nature genetics

Transposable elements have rewired the core regulatory network of human embryonic stem cells

Galih Kunarso[1,2], Na-Yu Chia[3,4], Justin Jeyakani[1], Catalina Hwang[1,5], Xinyi Lu[3,6], Yun-Shen Chan[3,7], Huck-Hui Ng[3,4,6–8] & Guillaume Bourque[1]

Detection of new genomic control elements is critical in understanding transcriptional regulatory networks in their entirety. We studied the genome-wide binding locations of three key regulatory proteins (POU5F1, also known as OCT4; NANOG; and CTCF) in human and mouse embryonic stem cells. In contrast to CTCF, we found that the binding profiles of OCT4 and NANOG are markedly different, with only ~5% of the regions being homologously occupied. We show that transposable elements contributed up to 25% of the bound sites in humans and mice and have wired new genes into the core regulatory network of embryonic stem cells. These data indicate that species-specific transposable elements have substantially altered the transcriptional circuitry of pluripotent stem cells.

Although it has been recognized that the gain and loss of regulatory elements are common features of eukaryotic genomes[1,2], most studies investigating this have been limited to the detection of binding events in one species followed by an *in silico* analysis of evolutionary conservation[3–5] or have been restricted by the scope and comparability of the functional datasets being analyzed[6,7]. To systematically explore the impact of newly arisen regulatory elements in a mammalian system, we generated matching datasets in human and mouse undifferentiated embryonic stem cells and studied the role of OCT4, NANOG and CTCF. The first two are known key regulators in embryonic stem cells[7,8], and the third is an important factor in the organization of regulatory blocks[9]. Previous studies have pointed to both similarities and differences between the expression profiles of these cells[10,11]. Additional insights gained about the evolution and the wiring of this core regulatory network could provide deeper understanding of pluripotent stem cells derived from various species[12].

We began our analysis by generating chromatin immunoprecipitation sequencing (ChIP-Seq) libraries for these three factors and then determined their genome-wide occupancy profile in human embryonic stem cells (see Online Methods). We used the full set of binding regions (**Fig. 1a**) to enable analyses of loci across a range of enrichment levels. Using these binding regions, a *de novo* motif-finding method recapitulated the known OCT4, NANOG and CTCF DNA binding motifs and helped confirm the quality of the data (Online Methods, **Supplementary Fig. 1** and **Supplementary Note**). Notably, the motifs defined from comparable mouse embryonic stem cell datasets[13] explained the human binding regions nearly as well (**Supplementary Fig. 1b**). This confirms the high similarity of the DNA-binding specificity of these proteins in human and mouse embryonic stem cells.

In a preliminary study, we suggested that the overlap between human and mouse binding regions in embryonic stem cells was limited[7]. However, that earlier assessment was hindered by the fact that the dataset of human samples was not genome wide and that the detection technologies used for each species were different (array based versus sequencing based). In contrast, the human datasets presented here enable a direct comparison to the mouse datasets previously obtained[13], and so, based on the regions detected in the human samples, we evaluated the proportion of regions that were also observed to be bound in mouse by looking for binding evidence within homologous windows of 1 kb in length (Online Methods). Overall, we found that only 2.0%, 1.9% and 16.7% of the regions occupied by OCT4, NANOG and CTCF in human were also occupied in mouse, respectively. Increasing the window sizes to 2 kb and 5 kb only had a moderate effect on the results (**Supplementary Fig. 2a**). Focusing on the top 10% most enriched regions, it is even more notable that only 3.8% of the OCT4 regions and 5.3% of the NANOG regions are conserved compared to 49.6% of the CTCF regions (**Fig. 1b**). To address potential issues with the sensitivity of the ChIP-Seq assays, we performed the converse analysis starting with the mouse binding regions and looking for evidence of binding in the human datasets but we also observed limited conservation (**Supplementary Fig. 2b**). Together, this confirms that the *in vivo* occupancy profiles of OCT4 and NANOG are notably different between human and mouse embryonic stem cells.

Recent studies have suggested that for a number of transcription factors, transposable elements have been a rich source of new binding sites[4,14]. We were interested in measuring whether this phenomenon was also a major contributing factor for the binding sites of OCT4,

[1]Computational and Mathematical Biology, Genome Institute of Singapore, Singapore, Singapore. [2]Duke–National University of Singapore Graduate Medical School, Singapore, Singapore. [3]Gene Regulation Laboratory, Genome Institute of Singapore, Singapore, Singapore. [4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [5]Princeton University, Princeton, New Jersey, USA. [6]Department of Biological Sciences, National University of Singapore, Singapore, Singapore. [7]Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore, Singapore. [8]Department of Biochemistry, National University of Singapore, Singapore, Singapore. Correspondence should be addressed to G.B. (bourque@gis.a-star.edu.sg).

Received 30 November 2009; accepted 20 April 2010; published online 6 June 2010; corrected online 13 June 2010; doi:10.1038/ng.600

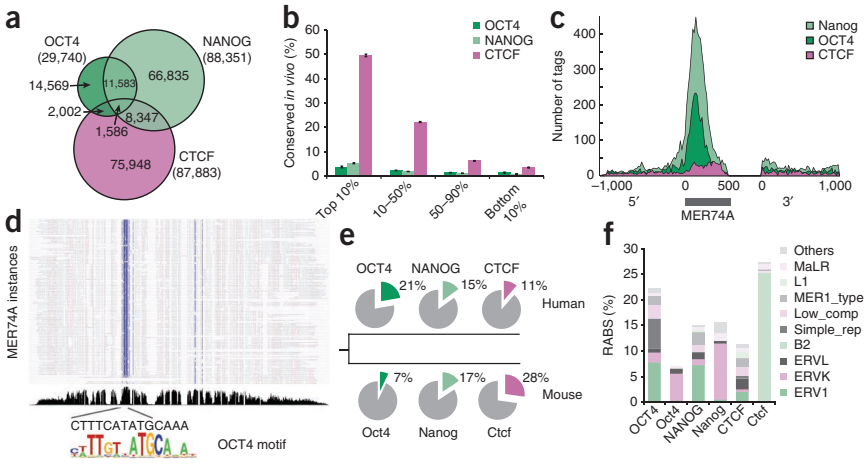

Figure 1 Genome-wide binding profiles of OCT4, NANOG and CTCF reveal limited evolutionary conservation and the role of transposable elements in facilitating binding site diversity. (**a**) Number of regions bound by OCT4, NANOG and CTCF in human embryonic stem cells. (**b**) Fraction of human binding regions for which the homologous region is also observed to be bound in mouse embryonic stem cells[13]. Human binding regions are split based on binding intensity into four groups. Error bars show 1 s.e.m. (**c**) Aggregate profile of mapped tags in and around the MER74A repeats from the ERV1 family. The bar under the graph highlights the location of the repeats and is flanked by 1-kb regions that are upstream (5′) and downstream (3′). (**d**) Multiple sequence alignment of the instances MER74A that are bound by OCT4. The graph displayed on the *x* axis shows the percent identity of each column in the alignment. Columns with more than 70% identity are in blue and highlight a region of higher sequence similarity. The consensus (ancestral) repeat sequence at that position corresponds well to the OCT4 binding motif. (**e**) Fraction of binding regions that correspond to repeat-associated binding sites (RABS) in human and mouse. (**f**) Contribution of the different families of repeats for each transcription factor in human and mouse.

NANOG and CTCF in human embryonic stem cells because this could affect the regulation of neighboring genes[15,16]. By calculating the observed overlap between the binding regions of each factor and the various repeat families, we were able to identify specific transcription factor-repeat associations that were more common than those expected by chance (Online Methods). For instance, even though there are only 767 LTR9B repeats from the endogenous retrovirus 1 (ERV1) repeat family in the human genome, we observed that 255 (33.2%) of these repeats are bound by OCT4. By chance, we would have only expected 3.1 (0.4%), and the number seen here corresponds to an 82-fold enrichment. We call such binding sites repeat-associated binding sites (RABS). Looking at the tag density in and around repeat instances of over-represented families, it is clear that specific regions of their ancestral sequence are preferentially targeted (**Fig. 1c** and **Supplementary Fig. 3**). Moreover, in many cases, aligning the bound instances of a given repeat family can show that the same region of the ancestral sequence has a high degree of sequence similarity among the bound sequences and harbors the cognate binding motif (**Fig. 1d**).

Collectively, we calculated that RABS accounted for 20.9%, 14.6% and 11.1% of the OCT4, NANOG and CTCF binding regions, respectively (**Fig. 1e** and **Supplementary Table 1**). Notably, the contributions of RABS were evenly distributed among the high- and the low-intensity binding regions for CTCF and were slightly skewed toward strongly bound sites for OCT4 and NANOG (**Supplementary Fig. 4**). For both OCT4 and NANOG, we found that the ERV1 repeat family is the largest contributor of RABS. In total, 2,464 (8.3%) of the OCT4 binding regions and 6,376 (7.2%) of the NANOG binding regions overlapped ERV1 repeats (**Fig. 1f**). Applying the same procedure to the mouse datasets showed that RABS accounts for 7.2%, 17.1% and 28.3% of the binding regions of Oct4, Nanog and Ctcf, respectively (**Fig. 1e**). It is notable that most of the families of transposable elements that have been exapted in the two species are different and correspond to species-specific sequences (**Fig. 1f** and **Supplementary Table 2**). Indeed, of the 6,231 OCT4 binding regions classified as RABS in human, only 58 (0.9%) have a homologous region in the mouse that is also bound.

To determine the functional relevance of RABS, we depleted human embryonic stem cells of *POU5F1* (also known as *OCT4*) by RNA interference (RNAi) and examined differential gene expression by microarray analysis. We processed the microarray data and identified 721 genes that were down regulated and 1,407 genes that were up regulated (Online Methods and **Supplementary Table 3**). When we checked whether the differentially expressed genes had binding within 20 kb of their transcription start site (TSS), we observed an enrichment of OCT4 and NANOG binding regions especially around the downregulated genes (Online Methods, **Fig. 2a** and **Supplementary Fig. 5a**). Moreover, we found that OCT4 regions overlapping a NANOG region were 1.85-fold over-represented in proximity to downregulated genes as compared to nonregulated genes (P value $< 1.0 \times 10^{-10}$, **Fig. 2b**). Similarly, conserved OCT4 regions were also enriched 1.96-fold ($P = 0.0011$). Also of note, OCT4 RABS showed a 1.86-fold enrichment ($P = 5.6 \times 10^{-8}$), and breaking up the RABS by repeat family revealed that the enrichment increased to 3.1-fold ($P = 2.5 \times 10^{-8}$) for binding

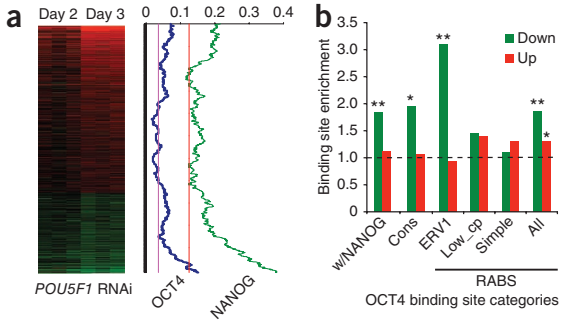

Figure 2 Binding sites and RABS are enriched in proximity to regulated genes. (**a**) Genes are rank-ordered by degree of induction (red) and repression (green) at 2 and 3 d after *POU5F1* RNAi treatment. The two plots on the right show the corresponding numbers (moving averages) of gene probes that have associated OCT4 or NANOG binding regions, respectively. Expected background levels are shown using a straight solid line. (**b**) Fold enrichment of different types of OCT4 binding regions in proximity of down- and upregulated genes relative to nonregulated genes. The categories of OCT4 binding regions are: overlapping a NANOG binding region (with NANOG), conserved *in vivo* (Cons) or classified as a RABS when overlapping an ERV1 repeat (ERV1), a low_complexity repeat (Low_cp), a simple_repeat (Simple) or any RABS family member (All).* $P < 0.01$, ** $P < 0.001$.

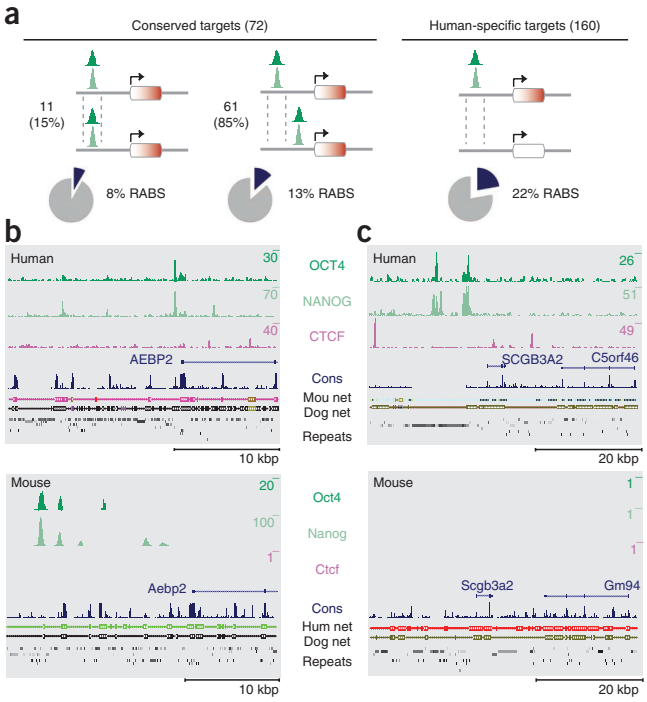

Figure 3 Binding profiles around regulated genes reveal functional binding site turnover and the presence of RABS in proximity of human-specific targets. (**a**) Genes that are downregulated after *POU5F1* RNAi treatment and that have an OCT4-NANOG binding region within 20 kb of their TSS are classified as conserved or human-specific targets according to their response in a similar experiment in mice. Conserved targets are further split into genes that have a conserved OCT4-NANOG binding region (left side) and genes that do not. Each human gene (top) is depicted with the mouse homologous gene (bottom) and is colored in red when it is differentially regulated following RNAi treatment. OCT4-NANOG binding sites are shown in green and aligned homologous regions are shown using gray dotted lines. Pie charts show the fraction of binding regions around the genes in a given category that correspond to RABS. (**b**) Although *AEBP2* is downregulated following *POU5F1* RNAi in both human and mouse, the binding profiles of OCT4 and NANOG around this gene are different (see **Supplementary Fig. 6a**). The size of the binding peaks are indicated on the right side of the tracks. Mouse or human (Mou/Hum Net) and dog (Dog Net) conservation tracks are displayed to show the conservation context of the promoters. (**c**) *SCGB3A2* is an example of a human-specific target and has two strong OCT4-NANOG binding regions in the promoter that are overlapping ERV1 repeat sequences that are absent in the mouse.

sites embedded in the ERV1 repeat family. This is strong evidence for a functional role of the OCT4-ERV1 sites in transcriptional regulation.

Given that the majority of the OCT4 and NANOG binding regions are different in humans and mice (**Fig. 1b**) and that we had access to matching *Pou5f1* RNAi data in mouse embryonic stem cells[7], we investigated the binding profiles around conserved gene targets in further depth. We compared the expression of orthologous genes between humans and mice and identified 137 genes that were downregulated in both human and mouse (conserved targets) following RNAi treatment (Online Methods and **Supplementary Table 4**). Included in this list is *POU5F1*, as well as a number of other factors implicated in embryonic stem cell biology (for example, *SOX2*, *NANOG*, *KLF4* and *DPPA4*). Although the strongest binding signal was observed in the immediate promoter of these genes, there was an enrichment of binding regions reaching up to 20 kb both upstream and downstream of the TSS (Online Methods and **Supplementary Fig. 5b**,**c**). In total, 72 of the 137 (53%) conserved targets had an OCT4-NANOG binding region, but only 11 of these (15%) were homologously bound in the mouse samples, whereas the other genes showed evidence of binding site turnover (**Fig. 3a** and **Supplementary Table 5**). For instance, *AEBP2*, which encodes a protein found in the PRC2 complex that is known to be important for stem cell self-renewal and differentiation[17], is a typical example of and shows evidence of binding site turnover (**Fig. 3b**). For this gene, the proximal promoter site in human overlaps a repeat that appears to be absent in mouse (**Supplementary Fig. 6a**). An exception to this is *SOX2*, which has a very well-conserved binding profile in mice and humans for the three factors considered here (**Supplementary Fig. 6b**).

Looking at the 584 genes that only showed downregulation in human embryonic stem cells, we found that 160 (27%) had an OCT4-NANOG binding region. Notably, for these human-specific targets, the fraction of binding regions corresponding to RABS was higher (22.5%) as compared to the conserved targets (12.4%). For instance, *SCGB3A2* (encoding secretoglobin, family 3A, member 2), which is downregulated following *POU5F1* RNAi treatment, contains two binding regions in its promoter that are bound by OCT4 and NANOG and that overlap ERV1 repeats (**Fig. 3c**). This gene, which was previously reported as one of the most highly expressed genes in human embryonic stem cells[18], is not regulated in mouse, but this difference can now be explained by the presence of

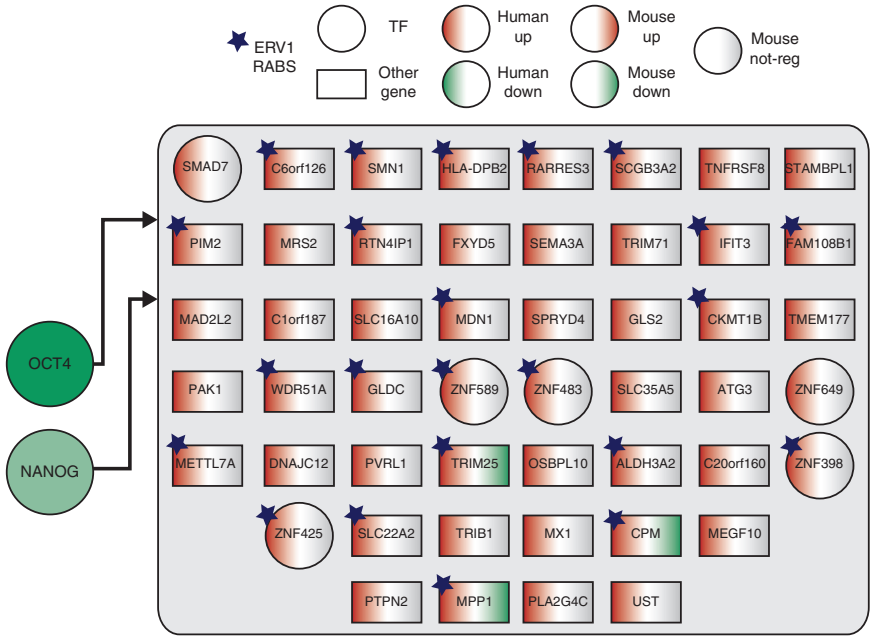

Figure 4 Transposable elements have wired new genes into the core regulatory network of human embryonic stem cells. Diagram showing the target genes that are regulated by OCT4 exclusively in human embryonic stem cells and that have a RABS within 20 kb of their TSS. Targets with an ERV1-RABS are highlighted with a blue star.

species-specific transposable elements. In total there are 50 human-specific targets that have a RABS, including 23 that have an ERV1-RABS (**Fig. 4**). We selected two of these ERV1-RABS and showed, using a luciferase assay, that they can drive enhancer activity and that this activity is ablated if the OCT4 motif is mutated (**Supplementary Note**). Together, these results suggest that many genes have been rewired into the core regulatory network of human embryonic stem cells following the insertion of transposable elements.

In summary, we found that CTCF has a stable occupancy profile not only across cell types[19] but also across species. In contrast, OCT4 and NANOG have very different binding profiles in human and mouse embryonic stem cells, with only ~5% of their sites being homologously occupied. The fact that there is also a limited concordance between regions experimentally observed to be bound and conserved elements, as determined from multispecies sequence alignments (**Supplementary Fig. 7**), implies that *in vivo* maps in the relevant species will be important in the study of many mammalian systems. Moreover, to help explain the vast occupancy differences, we showed that species-specific transposable elements have been an important source of new sites in both species. Using matched binding and expression datasets, we also demonstrated that many of these transposable element–derived sites are found in the vicinity of conserved target genes in human and mouse. Finally, beyond the genes that have similar expression profile changes in human and mouse, we were also able to identify a group of human-specific target genes that show evidence of having been added to the core regulatory network of human embryonic stem cells via the insertion of transposable elements. Although we do not expect all binding events to directly influence gene expression, this data adds important support to a seminal hypothesis on the impact of repeats on the evolution of transcription regulation[20–22]. Our results reveal the striking plasticity of the core regulatory network of mammalian embryonic stem cells and the importance that transposable elements have had in facilitating this functional turnover.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Raw sequence tags, peaks files and OCT4 RNAi expression files have been deposited to GEO with the accession code GSE21200.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the Agency for Science, Technology and Research (A*STAR) of Singapore.

AUTHOR CONTRIBUTIONS

H.-H.N. and G.B. designed the experiments. N.-Y.C., X.L. and Y.-S.C. performed the experiments. G.K. performed the data analysis with contributions from J.J. and C.H. G.B. wrote the manuscript with contributions from H.-H.N. and G.K.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Dermitzakis, E.T. & Clark, A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
2. Borneman, A.R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819 (2007).
3. Moses, A.M. *et al.* Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLOS Comput. Biol.* **2**, e130 (2006).
4. Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
5. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
6. Odom, D.T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732 (2007).
7. Loh, Y.H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
8. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
9. Bell, A.C., West, A.G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387–396 (1999).
10. Brons, I.G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
11. Tesar, P.J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
12. Rossant, J. Stem cells and early lineage development. *Cell* **132**, 527–531 (2008).
13. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
14. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. USA* **104**, 18613–18618 (2007).
15. Cohen, C.J., Lock, W.M. & Mager, D.L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105–114 (2009).
16. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
17. Cao, R. & Zhang, Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol. Cell* **15**, 57–67 (2004).
18. Sperger, J.M. *et al.* Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. *Proc. Natl. Acad. Sci. USA* **100**, 13350–13355 (2003).
19. Kim, T.H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
20. Davidson, E.H. & Britten, R.J. Regulation of gene expression: possible role of repetitive sequences. *Science* **204**, 1052–1059 (1979).
21. McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
22. Brosius, J. Retroposons—seeds of evolution. *Science* **251**, 753 (1991).

ONLINE METHODS

Whole-genome chromatin-immunoprecipitation datasets. The hESC line H1 (WA-01, passage 28)[23] was used for this study. The cells were cultured feeder free on Matrigel (Becton Dickinson)[24]. Condition medium used for culturing hESCs contained 20% knockout serum replacement, 1 mM L-glutamine, 1% nonessential amino acids, 0.1 mM 2-mercaptoethanol and an additional 8 ng/ml of basic fibroblast growth factor (Invitrogen) supplemented to the hESCs unconditioned medium. The medium was changed daily. The hESCs were subcultured with 1 mg/ml collagenase IV (Gibco) every 5–7 d. The H1 hESCs were cross-linked with 1% formaldehyde for 10 min at room temperature, and the formaldehyde was then inactivated by the addition of 125 mM glycine. ChIP-Seq was carried out as described previously[13]. Briefly, chromatin extracts containing DNA fragments with an average size of 500 base pairs (bp) were immunoprecipitated. Illumnia/Solexa adaptors were ligated to the ChIP DNA fragments (10 ng) and subjected to 15 cycles of PCR amplification. The fraction of fragments averaging 200 bp in length was selectively cut out from the gel and eluted by Qiagen gel extraction kit. Using the Illumina/Solexa platform, 13–22 million 36-bp tags were sequenced from these samples, out of which 9.6, 9.9 and 12.6 million tags were mapped uniquely to the human genome (NCBI36/hg18 assembly) using the ELAND program (see URLs). The antibodies used were Abcam (AB19857) for OCT4, R&D (AF1997) for NANOG and Upstate (07-729) for CTCF.

Finally, using the program MACS[25], the binding regions were ranked based on the enrichment of ChIP sequenced tags by comparing each ChIP library to an input library as a control. We identified 29,740, 88,351 and 87,883 peak regions for OCT4, NANOG and CTCF, respectively (**Supplementary Tables 6–8**). We defined the binding peaks as those above the *P* value cutoff of 1.00×10^{-5}. A number of factors will influence the resolution of the peak calling procedure (most notably initial fragment length and sequencing depth). For our analysis, we retained all peaks; however, it is possible that some of the peaks called in close proximity to each other might have originated from a single binding location.

***De novo* motif finding.** To find the motifs over-represented in the binding regions, we used the repeat-masked sequence from the regions 100 bp around the top 1,000 peaks of each transcription factor as input for the MDmodule program[26]. The highest-ranking motif in each library was similar to the known motif of the corresponding specific transcription factor (**Supplementary Fig. 1a**). For each identified motif, we scanned back the bound regions using a previously described method[7] with *e*-value cutoff of 0.001 to identify the binding peaks that had the motif (**Supplementary Fig. 1b**). We also did the same motif scan using the mouse PWMs previously identified[13] to calculate the proportion of human binding regions that can be explained by the mouse motif. Finally, scanning larger 600-bp windows centered around the middle of the bound regions revealed a strong enrichment for the recognition motifs especially within 60 bp of the peak (**Supplementary Fig. 1c**). Together these results help confirm the quality of the ChIP procedures.

Assessing conservation *in vivo* and *in silico*. To identify the binding regions conserved *in vivo*, we first extended each region identified in human to 50 bp, 200 bp or 1,000 bp (1 kbp) windows surrounding the peaks and used liftOver[27] with default parameters to determine the homologous regions on the mouse genome (NCBI36/mm8; **Supplementary Table 9**). For the rest of the study (to be conservative and to maximize overlap), we intersected the results from the 1-kbp windows with the mouse binding regions reported previously[13] to identify the conserved binding regions. We also did the converse, starting from the mouse binding regions. The *in vivo* conservation estimates obtained in this way could have been affected by the choice of antibodies in the two species, but it is encouraging to see in the human regions similar levels of enrichment for motifs obtained independently in human and mouse (**Supplementary Fig. 1b**). This helps confirm the high similarity of the DNA binding specificity for these proteins in the two species and supports the comparability of the datasets. Finally, for the *in silico* analysis, we identified the human binding regions that overlap the 28-Way PhastCons Elements track[28] from the UCSC Genome Browser[27] using centered windows of fixed length (50 bp, 100 bp and 200 bp). The results are shown in **Supplementary Table 10**.

Identification of RABS. We used the 200-bp window surrounding the center of the transcription factor binding regions and intersected these with the RepeatMasker (see URLs) track from UCSC Genome Browser to find the number of overlaps of each transcription factor's binding regions with specific repeats. We also annotated each binding region with respect to its nearest RefSeq genes, up to a 100-kbp distance. We separated the binding regions into six categories according to the peak location: TSS (within 1 kbp of a TSS), promoter (up to 5 kbp upstream of TSS), intragenic (within the RefSeq gene boundary), proximal (up to 10 kbp away from the gene boundaries), distal (up to 100 kbp away from the gene boundaries) and desert (more than 100 kbp away from any RefSeq genes). Next, we generated a random dataset of 200,000 regions with the same annotation distribution as the true regions and intersected with the RepeatMasker track to obtain the expected number of overlaps of each transcription factor with repeat elements. We then used a one-sided binomial test to compare the observed number of repeats intersecting the true binding regions with the expected numbers from the annotation-matched background. We identified RABS as those repeats with statistically significant ($P < 1 \times 10^{-5}$) association with a transcription factor's binding regions. We also did the same analysis for the mouse binding regions.

Microarray expression analysis, target identification and network analysis. The background-adjusted Illumina results were normalized using MeV by performing $\log_2$ transformation, followed by median centering on samples and median centering of genes across the samples. We used SAM[29] with 5% false discovery rate and >1.5-fold cutoff to find the genes with statistically significant changes in expression upon RNAi treatment. We noted that depletion of *POU5F1* by RNAi induced rapid differentiation of human embryonic stem cells. Therefore, the gene expression profile is a combination of primary and secondary gene expression changes. For the mouse RNAi results, we used the data as previously provided[7]. To determine an appropriate distance cutoff to associate binding regions to genes, we looked at the absolute enrichment of OCT4-NANOG binding regions in proximity of downregulated RefSeq genes (**Supplementary Fig. 5a**). To maximize enrichment and comprehensiveness but also limit the level of background noise, we identified targets of each transcription factor in each genome as genes with binding regions within 20 kbp of its TSS. We sorted the expression changes of the genes and to display general binding patterns we used a sliding window one-eighth the size of the gene list and calculated the proportion of the changing genes that are bound by each transcription factor. We compared this proportion with the number of genes in the whole array that were bound by the transcription factor as the background. *P* values associated with fold enrichments were calculated using a one-sided binomial proportion test.

Finally, to identify homologous genes in human and mouse, we selected the longest transcript to represent each RefSeq in each species and used liftOver to convert the coordinates into the other species. We then intersected the new coordinates with the RefSeq genes of that particular genome and identified those genes that intersect in the same strand as the homologous gene pairs from the two species.

URLs. ELAND, http://www.illumina.com/software/; RepeatMasker, http://www.repeatmasker.org/.

23. Thomson, J.A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
24. Xu, C. *et al.* Feeder-free growth of undifferentiated human embryonic stem cells. *Nat. Biotechnol.* **19**, 971–974 (2001).
25. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
26. Conlon, E.M., Liu, X.S., Lieb, J.D. & Liu, J.S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* **100**, 3339–3344 (2003).
27. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
28. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
29. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).