Stem Cell Reports
Resource
OPEN ACCESS

Directed Evolution of Reprogramming Factors by Cell Selection and Sequencing

Veeramohan Veerapandian,[1,2,3] Jan Ole Ackermann,[1,2] Yogesh Srivastava,[1,2,3] Vikas Malik,[1,2,3] Mingxi Weng,[1,2,4] Xiaoxiao Yang,[1,2] and Ralf Jauch[1,2,4,*]

[1]CAS Key Laboratory of Regenerative Biology, Joint School of Life Sciences, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, and Guangzhou Medical University, Guangzhou, China
[2]Genome Regulation Laboratory, Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China
[3]University of Chinese Academy of Sciences, No.19A Yuquanlu, Beijing, China
[4]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, L4-41, Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong SAR, China
*Correspondence: ralf@hku.hk
https://doi.org/10.1016/j.stemcr.2018.07.002

SUMMARY

Directed biomolecular evolution is widely used to tailor and enhance enzymes, fluorescent proteins, and antibodies but has hitherto not been applied in the reprogramming of mammalian cells. Here, we describe a method termed directed evolution of reprogramming factors by cell selection and sequencing (DERBY-seq) to identify artificially enhanced and evolved reprogramming transcription factors. DERBY-seq entails pooled screens with libraries of positionally randomised genes, cell selection based on phenotypic readouts, and genotyping by amplicon sequencing for candidate identification. We benchmark this approach using pluripotency reprogramming with libraries based on the reprogramming factor SOX2 and the reprogramming incompetent endodermal factor SOX17. We identified several SOX2 variants outperforming the wild-type protein in three- and four-factor cocktails. The most effective variants were discovered from the SOX17 library, demonstrating that this factor can be converted into a highly potent inducer of pluripotency with a range of diverse modifications. We propose DERBY-seq as a broad-based approach to discover reprogramming factors for any donor/target cell combination applicable to direct lineage reprogramming *in vitro* and *in vivo*.

INTRODUCTION

The forcible expression of defined transcription factor (TF) cocktails can effectively engage and reprogram the epigenome of somatic cells, leading to drastic cell-fate conversions. A four-factor TF cocktail consisting of SOX2, OCT4, KLF4, and C-MYC (SOKM) directs pluripotency reprogramming in mouse and human cells (Takahashi et al., 2007; Takahashi and Yamanaka, 2006). Alternative TF cocktails have been shown to directly interconvert somatic cell types, a process termed direct lineage reprogramming, bypassing the intermediate step of pluripotency (Graf and Enver, 2009; Tanabe et al., 2015). However, the rate, quantity, reproducibility, and quality of cells produced by reprogramming technologies are often poor and pose challenges to translate this method for routine clinical diagnostics or cell-based therapies. For example, in pluripotency reprogramming under serum/leukemia inhibitory factor (LIF) conditions, less than 0.1% of mouse embryonic fibroblasts (MEFs) give rise to induced pluripotent stem cells (iPSCs). Only a select subset of TFs is capable of directing cell-fate conversions, and the unique molecular properties endowing them with the competence to reprogram are only poorly understood. Highly homologous TFs function differently in pluripotency reprogramming experiments. For example, if SOX17 replaces SOX2 or if OCT6 replaces OCT4, the reprogramming activity of four-factor cocktails is lost (Jauch et al., 2011; Nakagawa et al., 2008). Intriguingly, the uniqueness of reprogramming TFs appears to rely on subtle molecular features. Rationally introduced point mutations that direct the DNA-dependent dimerization can convert SOX17 and OCT6 into pluripotency reprogramming factors, although the wild-type proteins induce endodermal or ectodermal cell lineages, respectively (Aksoy et al., 2013b; Jauch et al., 2011; Jerabek et al., 2017). Apparently, protein engineering of endogenous factors can profoundly switch and enhance the function of reprogramming TFs. As our understanding of sequence-function relationships in transcriptional control is incomplete, rational design approaches suffer from major limitations. We therefore decided to let directed evolution take care of this problem. Directed evolution is commonly used to install new qualities to enzymes, fluorescent proteins, receptor-ligand pairs, or antibodies (Arnold, 2015). Here we asked whether we could use the phenotypes of mammalian cells to select for artificially improved proteins. We report directed evolution of reprogramming factors by cell selection and sequencing (DERBY-seq) combining cellular reprogramming with pooled libraries,

isolation of cells based on desired phenotypes, and amplicon sequencing for variant detection.

RESULTS

Design and Pooled Screening with Randomized eSOX Libraries

To benchmark DERBY-seq, we selected the reprogramming of MEFs to iPSCs as phenotypic readout. To design DERBY-seq libraries, we followed two guiding principles. First, amino acids should map to structurally characterized domains with a potential to either directly influence the interactions with DNA and chromatin or to modify the association with partner factors. Second, amino acids should be conserved among orthologs but show some variability among paralogs. We reasoned that such amino acids are good candidates to confer specific functions in transcriptional control. In the case of KLF4, structures for binary protein-DNA complexes are available but the structural basis for protein-protein interactions in the context of regulatory DNA is unknown (Schuetz et al., 2011). Thus, for KLF4 the randomization of amino acids involved in the sequence-specific base readout could be selected for library design (Figures S1A and S1B). SOX2 and OCT4 have been structurally analyzed in a number of different configurations that include heterodimeric and homodimeric complexes on different composite DNA elements (reviewed in Hou et al., 2017; Malik et al., 2018). In consequence, structural information is available for protein-DNA as well as protein-protein contact interfaces. *Sox* genes possess a 79-amino-acid high-mobility group (HMG) box enabling binding to the minor groove of the DNA with sequence specificity. Besides DNA recognition, the HMG box also facilitates the interaction with protein partners in a context-dependent manner. We thus selected the structural scaffold of the HMG box to establish the DERBY-seq method. To generate artificially evolving SOX (eSOX) libraries, we selected three residues of helix 3 in the HMG box domain that are variable among the 20 paralogous SOX factors encoded in mouse or human genomes and play a role in the DNA-dependent dimerization with OCT4 (Jauch et al., 2011; Merino et al., 2014; Ng et al., 2012; Remenyi et al., 2001) (Figures 1A and 1B). NNK sequence diversification was used to cover all 20 amino acids with 32 codons (Figure S1C) (Packer and Liu, 2015). In this way, we randomized E46, I53, and K57 in SOX2 and the homologous L46, V53, and E57 in SOX17 (HMG box numbering convention; Bowles et al., 2000), leading to libraries with 20^3 = 8,000 variants excluding truncations caused by the single remaining STOP codon. Randomizing four amino acid residues would lead to substantially larger 20^4 = 160,000 variant libraries. As we aspired to probe the reprogramming activity of the whole sequence space of the eSOX libraries, we opted for the 8,000 variant libraries for our experiments. To establish our pooled library screens, we used the reprogramming of MEFs carrying a GFP transgene controlled by regulatory sequences of *Oct4* permitting the identification of pluripotent cells (Figure 1C). Libraries were prepared as retroviral mixtures and used to transduce MEFs in four-factor combination (4F: *Oct4*, *Klf4*, and *c-Myc* [OKM] + *eSox*) or three-factor combination (3F: *Oct4* and *Klf4* [OK] + *eSox*) under LIF/serum/vitamin C conditions (Esteban et al., 2010) (Figures 1C, S1D, and S1E). Under these conditions, SOX2-containing cocktails can direct pluripotency reprogramming and typically yield 50–100 GFP-positive colonies per well of a 12-well plate by day 12 while the replacement of SOX2 with SOX17 impairs the capacity of 3F and 4F cocktail to generate iPSCs (Figures 2A and S1F). However, cocktails in which SOX2 was replaced with eSOX2 or eSOX17 libraries yielded a high quantity of GFP-positive colonies, demonstrating that pooled screens with randomized factors are feasible and allow for the separation of reprogramming competent and incompetent variants (Figures 2A, 2B, S1F, and S1G). Surprisingly, the cocktails containing randomized libraries outperformed wild-type SOX2; in particular, the library based on the otherwise inactive SOX17 shows elevated colony numbers and a higher yield of GFP-positive cells (Figures 2A, 2B, S1F, and S1G).

Identification and Selection of Variants from Pooled Screens

We next performed preparative experiments with eSOX2 and eSOX17 libraries under 3F and 4F conditions in three independent biological experiments with three technical replicates each in 6-well plates. At reprogramming days 12–14, cells were trypsinized and single-cell suspensions containing heterogeneous populations of GFP-positive and GFP-negative cells were separated by fluorescence-activated cell sorting (FACS; Figures S2A–S2C). We observed an increased proliferation rate of cells transfected with eSOX libraries and SOX17 as compared with SOX2 (Figure S2D). To genotype candidates from eSOX libraries, we amplified transgenes from genomic DNA in a first round of PCR with primer pairs specifically amplifying exogenously provided *Sox* factors. In the subsequent PCR cycles, sequencing adaptors and barcodes were added (Figures S2E and S2F). Each library was sequenced in two technical replicates. Deep sequencing generated ~0.5 million raw reads per sample. Randomized codons were translated and tripeptide occurrences counted. To probe for PCR bias, we sequenced the input library in two technical replicates, and used two different numbers of PCR cycles and four dilutions (Figure S2E). We observed a high correlation between all control reactions, suggesting that imbalances in read counts

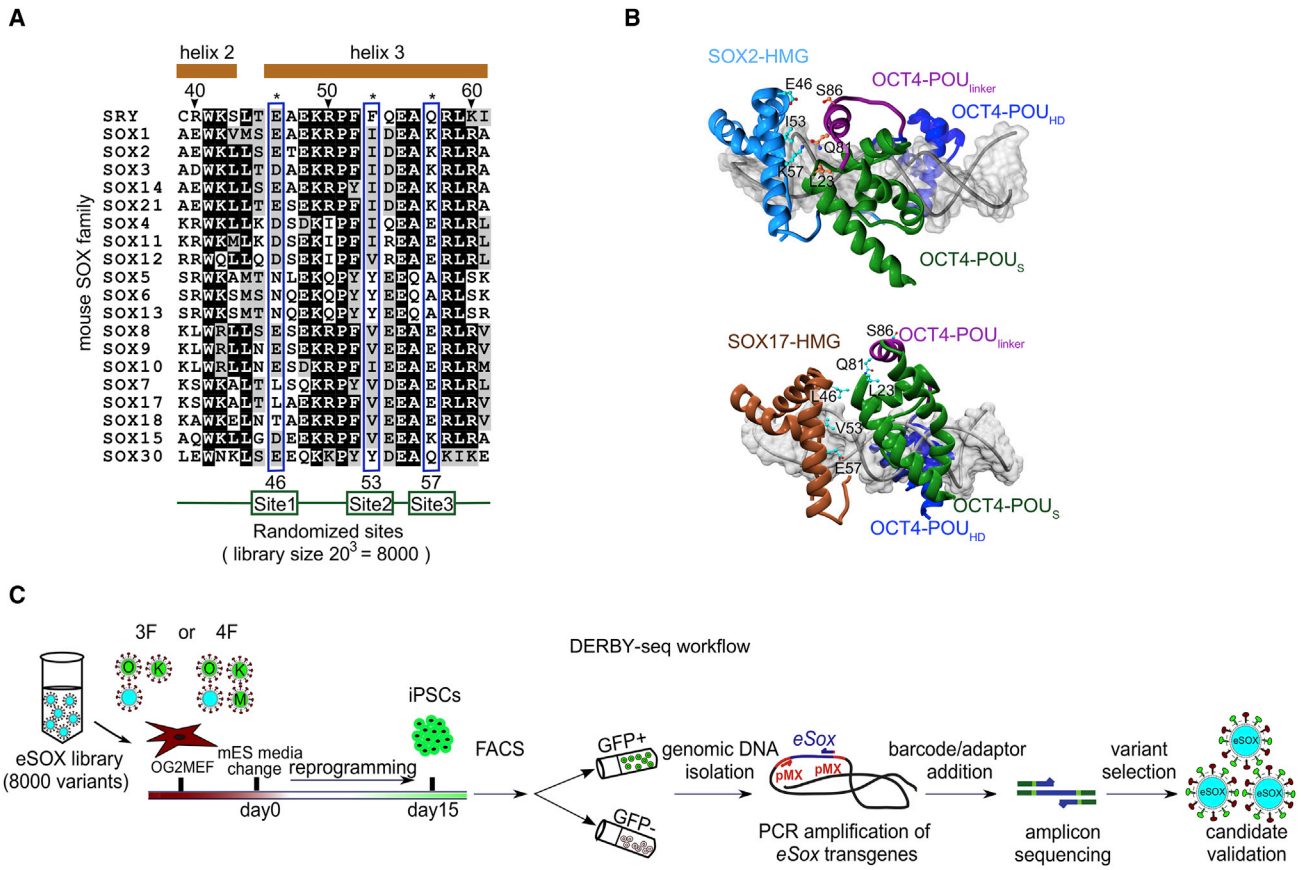


Figure 1. Design of DERBY-Seq Libraries and Experimental Strategy

(A) Multiple sequence alignment of the portion encompassing helix 3 of the high-mobility group (HMG) box of 20 paralogous mouse SOX proteins. Helix 3 mediates DNA-dependent dimerization with OCT4 on canonical and compressed DNA elements with juxtaposed *Sox* and *Oct* half-sites. The boxes mark sites 1, 2, and 3 and correspond to E46/I53/K57 for SOX2 and L46/V53/E57 for SOX17 subjected to randomization with NNK codons (Figure S1C).

(B) Structural models of the SOX2-HMG/OCT4-POU dimers on canonical *SoxOct* DNA elements and of the SOX17-HMG/OCT4-POU dimers on compressed DNA elements. Residues mediating the DNA-dependent heterodimer formation are labeled and shown as ball-and-sticks. Structural cartoons were prepared using Chimera (https://www.cgl.ucsf.edu/chimera/).

(C) Schematic representation of the DERBY-seq workflow. A pooled library of 8,000 eSOX variants was used in three biological replicates to reprogram 90,000 OG2-MEFs (30,000 MEFs plated per well of a 6-well plate) to iPSCs in LIF/serum/vitamin C medium using 3F (eSOX library plus OK) or 4F (eSOX library plus OKM) conditions. After FACS, the genomic DNA is isolated and fragments encompassing randomized codons are amplified in a two-step (eSOX17) or three-step (eSOX2) PCR procedure, and submitted for amplicon sequencing (Figures S2E and S2F).

from sorted cells arose for biological rather than technical reasons (Figures S2G and S2H). Read counts in GFP-positive and GFP-negative cell populations in biological replicates correlate better for 3F than 4F experiments, presumably because of high proliferation rates and transcriptional noise introduced by *c-Myc* (Figures S2D and S2I). We scored for differentially enriched reads of eSOX variants in GFP-positive and -negative cells using $\log_2$ fold change and p_{adj} (adjusted p value) scores determined by DESeq2 (Love et al., 2014) (Figures 3A and S2J; Tables S2 and S3). We chose high-ranking variants from the DESeq2 analysis (based on base mean, $\log_2$ fold change, and p_{adj}, Tables S2 and S3) to select candidates for validation and also took the identity of affected amino acids into account.

DERBY-Seq Identifies Functionally Enhanced Reprogramming TFs

We next prepared retroviruses of individual mutants identified in our eSOX2 screens and tested their capacity to induce pluripotency in comparison with their wild-type counterparts (Figures 3B–3D, S2K, and S2L). A $eSOX2^{NRR}$ variant (where NRR refers to the SOX2 E46N, I53R, K57R triple mutant) reproducibly outperformed wild-type SOX2 (Figures 3B, 3C, and 3E). Likewise, some candidates

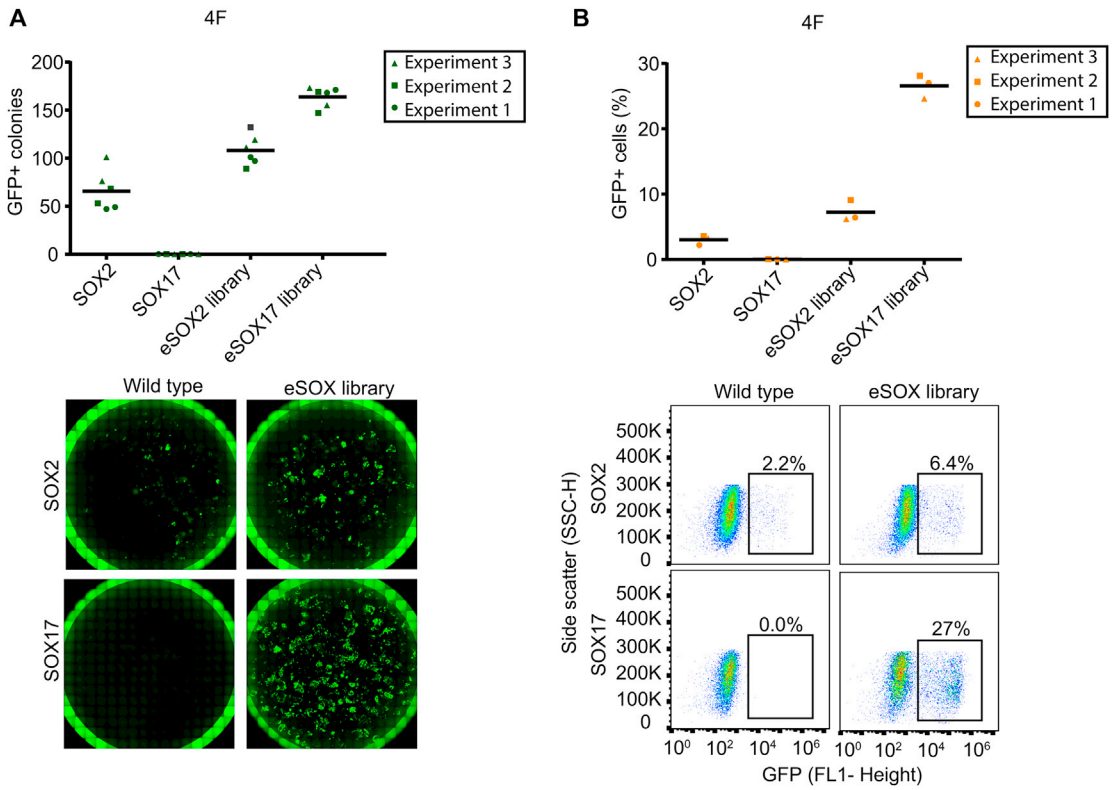


Figure 2. eSOX Libraries Effectively Induce Pluripotent Stem Cells

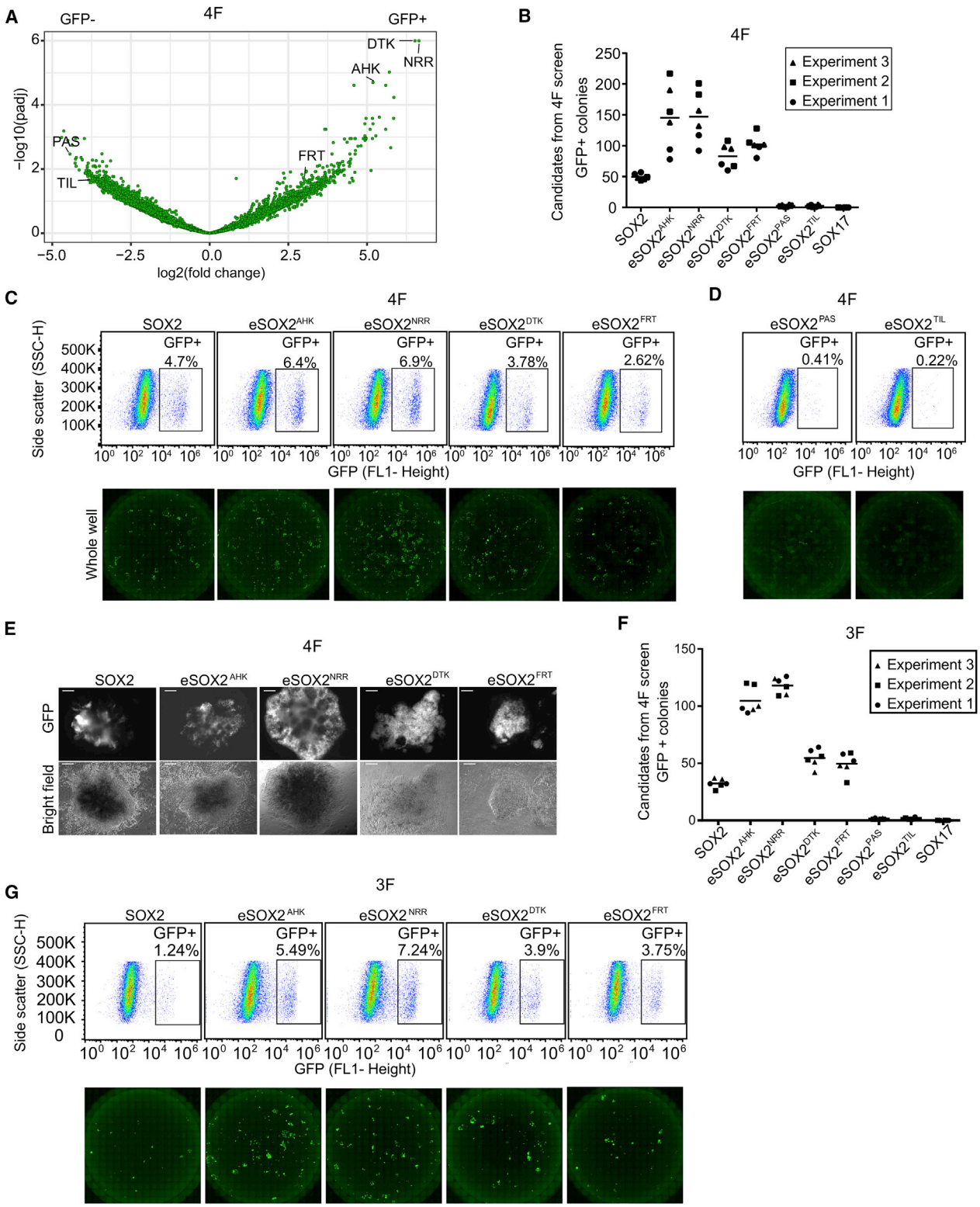
(A) The upper panel shows the counts of GFP-positive iPSC colonies from three independent biological experiments performed in technical duplicates; the black bar indicates the mean. The lower panel shows representative whole-well scans (from 12-well plates) of eSOX2 and eSOX17 libraries compared with wild-type SOX2 and SOX17 controls at day 12 of reprogramming for 4F conditions.

(B) The upper panel shows the percentages of GFP-positive cells after FACS analysis at day 12 performed in three biological replicates; the black bar is the mean. The lower panel shows representative FACS plots to illustrate the gating strategy for analytical experiments with pMX-GFP and pMX-Sox17 controls for 4F condition.

from the 3F eSOX2 screen also outperformed wild-type SOX2, similar to 4F candidates (Figures S2K, S2L, and S3I). To validate the robustness of our screen, we also tested a number of variants identified in GFP-negative populations in addition to GFP-positive variants and found that the majority indeed disrupt the reprogramming activity of SOX2 (Figures 3D and S2L). We found that the reprogramming enhancement of candidates identified under 4F conditions is further accentuated under 3F conditions (Figures 3F and 3G). Moreover, when introduced into equivalent positions in SOX17, NRR and AHK tripeptides convert SOX17 into a potent inducer of pluripotency (Figure 4).

DERBY-Seq Identifies Variants that Convert SOX17 into a High-Performance Pluripotency Inducer

We next tested candidates derived from the library of the otherwise reprogramming incompetent SOX17 as scaffold (Figures 4A and 4B). We selected a total of 22 (4F screen) and 17 (3F screen; Table S4) eSOX17 variants significantly enriched in GFP-positive cell populations for validation experiments. All of these eSOX17 variants exhibited potent pluripotency reprogramming activity (Figures 4C–4F, S3A, and S3B). Pluripotency-promoting tripeptides derived from the eSOX17 libraries are highly diverse and include WHC (where WHC refers to a SOX17 L46W, V53H, E57C triple mutant and analogous abbreviations are used for other variants), FNV, SLQ, DYC, or HQK variants, respectively. Of note, a validation experiment revealed a number of eSOX17 variants capable of inducing pluripotency although they were enriched in GFP-negative cell populations (Figures S3C and S3D). Apparently, an overwhelming number of eSOX17 variants in the starting library possess the competency to induce pluripotency. This is consistent with the highly efficient pluripotency reprogramming activity of the eSOX17 starting library itself (Figures 2, S1F, and S1G). Inactive eSOX17 variants identified in 3F and 4F conditions often retained the glutamate at position 57 (the last residue of the L46V53E57 tripeptide of the SOX17 HMG box), indicating that this amino acid constitutes a major barrier blocking



(legend on next page)

the pluripotency reprogramming activity of the native SOX17 protein (Figures S3E and S3F). Conservative replacements of E57 by aspartate or glutamine could also impair reprogramming in the context of some of the tripeptides (Figure S3F). To further dissect the roles of the three amino acids at positions 46, 53, and 57, we selected the high-performance eSOX17^{FNV} variant and individually tested which of the three mutations is most critical. We found that the eSOX17^{E57V} mutation is necessary and sufficient to convert SOX17 into a high-performance pluripotency reprogramming factor (Figure S3G). To evaluate and compare reprogramming efficiencies, we selected different time points for 3F, 4F, eSOX2, and eSOX17 conditions to reliably count colonies derived from independent reprogramming events and before merging of adjacent colonies (Figures S3H–S3J). That is, for highly efficient variants and conditions we counted earlier than for conditions with moderate efficiency. We selected the high-performance eSox variants eSOX2^{NRR}, eSOX17^{FNV}, and eSOX17^{WHC} to establish at least two clonal lines for each of them for characterization experiments. Pluripotent colonies derived from eSOX variants maintained good cell morphologies after 4–5 passages (Figures S4A and S4B), expressed critical pluripotency markers while silencing transgenes (Figures 4G and S4C–S4E), maintained a normal karyotype (Figure S4F), and exhibited a global gene expression profile reminiscent of embryonic stem cells (ESCs) (Figure S4G). This indicates that artificial factor evolution does not compromise the quality of the reprogrammed cells.

eSOX Variants Accelerate the Activation of the Pluripotency Network

We next sought to study the mechanism underlying the enhanced reprogramming activity of the artificially evolved SOX factors. Thus, we performed RNA sequencing (RNA-seq) using bulk cell populations collected at reprogramming days 3, 6, and 9 under 4F conditions for the eSOX variants eSOX2^{NRR} and eSOX17^{FNV} as well as SOX17, SOX2, and GFP controls. We also sequenced two iPSC lines for the eSOX variants eSOX2^{NRR}, eSOX17^{FNV}, and eSOX17^{WHC} (Figure 5A). Principal component analysis indicates similar reprogramming trajectories for SOX2, eSOX2^{NRR}, and eSOX17^{FNV} (Figures 5B and S5A). However, cells transfected with either eSOX2^{NRR} or eSOX17^{FNV} alongside OKM acquire expression profiles characteristic for cells transiting to a pluripotent state substantially earlier than SOX2-expressing cells. For example, the expression profile of SOX2 at day 6 resembles that of eSOX17^{FNV} at day 3, and the expression profile of eSOX17^{FNV} at day 6 is more advanced than the SOX2 expression profile at day 9 (Figures 5B and S5A). Consistently, expression of early and late pluripotency markers such as *Dppa3*, *Esrrb*, *Nanog*, *Prdm14*, *Utf1*, *Lin28a*, *Dppa5a*, *Dnmt3l*, *Sall4*, and *Zic3* is activated earlier and more strongly in eSOX17^{FNV} or eSOX2^{NRR} conditions as compared with wild-type SOX2 (Figures 5C, S5B, and S5C). We conclude that cells reprogrammed with eSOX factors do not take an alternative route to transit from a somatic to a pluripotent state. Alternative routes were, for example, reported for cells reprogrammed with chemical cocktails (Zhao et al., 2015). Rather, eSOX2^{NRR} and in particular eSOX17^{FNV} outperform wild-type SOX2 by activating the pluripotency network faster and in a higher proportion of cells.

Expression programs elicited by the reprogramming of incompetent SOX17-OKM and GFP-OKM cocktails are not identical (Figures S5A and S6A). In GFP-OKM conditions, cells progress to a global expression signature resembling day 3 (eSOX17^{FNV} and eSOX2^{NRR}) or day 6 (SOX2), suggesting that this cocktail initiates reprogramming but quickly derails without activating the pluripotency network (Figures 5B and S5A). In SOX17-OKM conditions cells appear to enter an alternative route (Figures 5B and S5A [rightmost panel]). Inspection of differentially expressed genes at day 9 in SOX17 and eSOX17^{FNV} conditions revealed that SOX17-expressing cells show an elevated expression for a number of somatic genes normally expressed in the cardiovascular system and during muscle development, including *Mef2c*, *Tie1*, *Robo4*, and *Gfap* (Figures 5D, 5E, and S5B). This suggests that under

Figure 3. DERBY-Seq Identifies Artificially Evolved and Enhanced SOX Factors

(A) Volcano plot showing the differential enrichment of eSOX2 variants in 4F condition. Each dot represents a variant. eSOX2 variants selected for validation experiments are marked.

(B) Colony count data from validation experiments using variants selected from the eSOX2 4F screen and Sox2 and Sox17 wild-type controls performed in biological triplicates each performed in technical duplicates. The black bar indicates the mean.

(C and D) The upper panel depicts FACS plots and the lower panel whole-well scans using a GFP fluorescence channel. Selected variants identified in GFP-positive cells (C) and GFP-negative cells (D) were chosen to be tested for GFP activity under 4F condition.

(E) Representative images of iPSC colonies generated by eSOX2 candidates identified in 4F screens. Scale bars, 100 μm.

(F) Colony count of candidates selected from the 4F screen but tested in 3F condition. Experiments were performed in biological triplicates and technical duplicates. The black bar indicates the mean.

(G) FACS plots and whole-well scans of candidates from the 4F screen tested under 3F condition observed at day 12.

See also Figures S1 and S2.

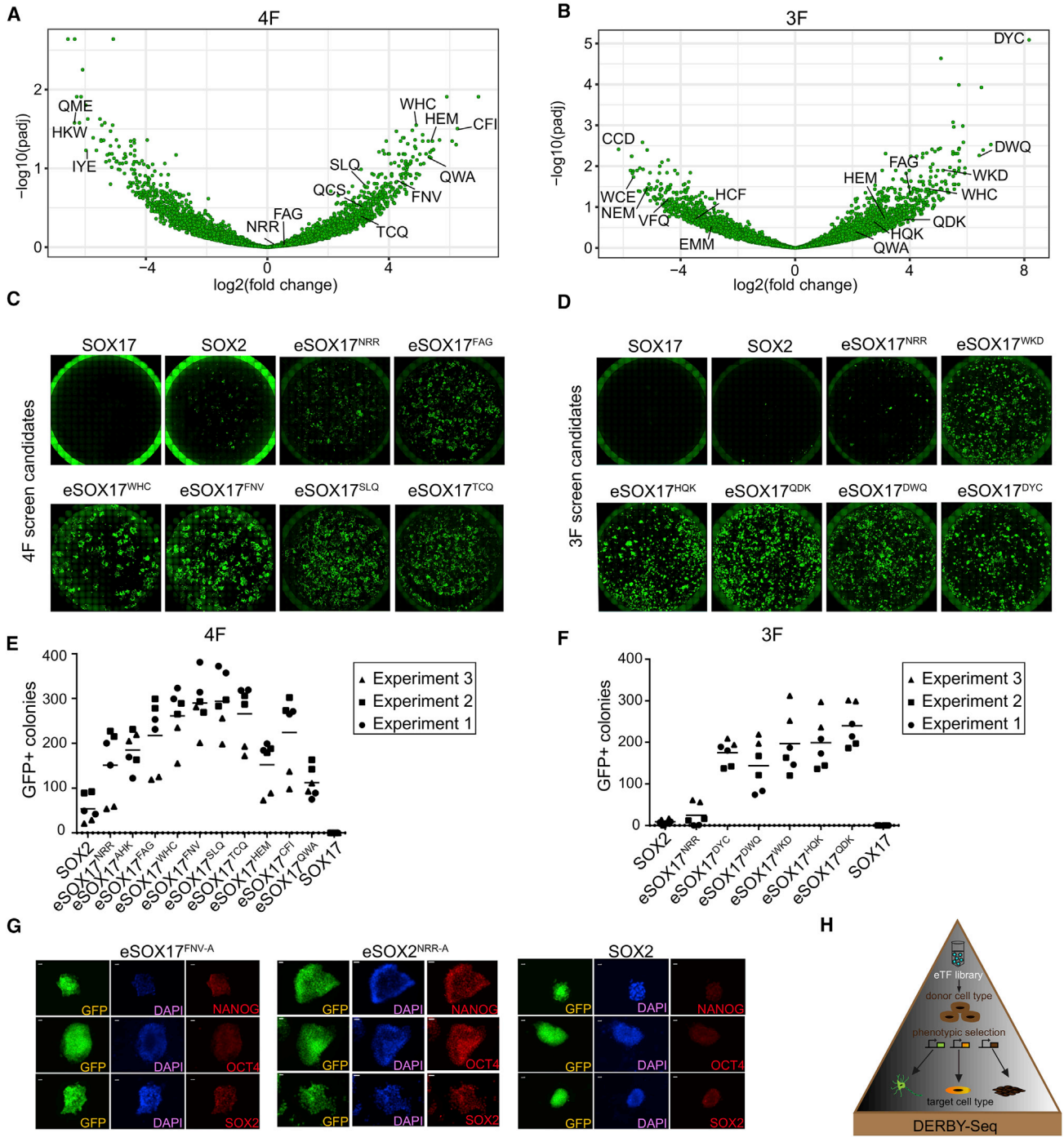


Figure 4. DERBY-Seq Identifies Variants that Convert SOX17 into a Potent Pluripotency Inducer

(A and B) Volcano plots showing the differential enrichment of eSOX17 variants in 4F (A) and 3F (B) conditions. Every dot represents an eSOX variant. Selected tripeptides are marked.

(C and D) Whole-well scans for validation experiments for eSOX17 variants selected from screens performed in 4F (C) or 3F (D) conditions.

(E and F) Colony count data for eSOX17 variants identified in 4F (E) and 3F (F) screens. Experiments were performed in biological triplicates and technical duplicates. The black bar indicates the mean. Colonies were counted at day 8 (4F) or day 10 (3F). See also Figure S3.

(legend continued on next page)

pluripotency reprogramming conditions SOX17 is directed to a set of target genes it would normally regulate in development, such as those required for the specification of the cardiac mesoderm (Liu et al., 2007).

We next asked how the mutations to SOX17 affect its genomic binding profile. To examine this, we performed chromatin immunoprecipitation sequencing (ChIP-seq) using antibodies for SOX17 at days 3 and 6 for SOX17-OKM and eSOX17^{FNV}-OKM conditions (Figure 5A). At both time points SOX17 and eSOX17^{FNV} exhibit profoundly different binding profiles and only a small fraction of sites is occupied by both factors (Figures S6B and S6C). *De novo* motif analysis using homer (http://homer.ucsd.edu/homer/motif/) revealed the canonical *SoxOct* DNA element as the top-scoring motif for eSOX17^{FNV} at days 3 and 6 (Figure 6A). The canonical motif consists of a CATTGTT-like *Sox* element juxtaposed to an ATG CAAAT-like octamer element and is found in the enhancers of many pluripotency genes (Chen et al., 2008; Knaupp et al., 2017; Whyte et al., 2013). In SOX17 conditions, however, the preference for the canonical motif is not observed. At day 3, a single *Sox* element is the top-scoring motif. At day 6 an alternative compressed version of a composite *SoxOct* element is most strongly enriched (Figure 6A). In the compressed *SoxOct* motif one base pair is eliminated, bringing the *Sox* and *Oct* half-sites closer together. We have previously found that heterodimeric binding of SOX17/OCT4 at compressed DNA elements directs the specification of the extraembryonic endoderm (Aksoy et al., 2013a). Only SOX17, but not SOX2, can form DNA-dependent heterodimers with OCT4 on compressed DNA elements while SOX2 more effectively than SOX17 associates with OCT4 on canonical DNA elements (Jauch et al., 2011; Merino et al., 2014; Ng et al., 2012). We next defined eSOX17^{FNV}-bound locations containing canonical *SoxOct* elements at days 3 or 6 (Figures 6B–6D). Wild-type SOX17 is unable to target these sites. On the contrary, in particular at day 6, SOX2 and OCT4 occupy these locations (Knaupp et al., 2017). Conversely, locations with matches to the compressed motif bound by SOX17 at days 3 or 6 are devoid of ChIP-seq signals for eSOX17^{FNV}, SOX2, and OCT4. This drastic change in binding profiles is illustrated by sites with canonical motifs in pluripotency super-enhancers (Whyte et al., 2013) near *Klf13* and *Sox2* and sites with compressed motifs near *Smad2* and *Id2* (Figures 6C and 6D). We conclude that the switch in genomic binding, transcriptional regulation, and pluripotency reprogramming function of eSOX variants is tied to their enhanced targeting of pluripotency enhancers in direct partnership with OCT4.

DISCUSSION

SOX factors are versatile regulators of cellular identity. Individual SOX TFs regulate different genes in the context of different cells leading to alternative phenotypic consequences. Family members largely fulfill non-redundant roles despite strong sequence conservation in functional domains and near-identical preferences for DNA *in vitro*. Therefore, the functional uniqueness and context specificity are likely caused by subtle sequence variations at critical molecular interfaces that determine chromatin engagement, with profound consequences on gene expression programs and cellular fate decisions. The DERBY-seq method enables the study of sequence-function relationships of proteins regulating gene expression and mammalian cell fates, analogously to the deep mutational scanning approach pioneered in phage display assays and using yeast as functional readout (Fowler et al., 2010; Fowler and Fields, 2014). Deep mutational scanning studies of a WW domain using phage display showed that the majority of mutations to a protein are deleterious and mitigate or destroy its capacity to bind natural ligands (Fowler et al., 2010). Here we find that many mutations to SOX17 produce a highly potent inducer of pluripotency, although the wild-type protein is unable to direct pluripotency reprogramming. Wild-type SOX2 and SOX17 are driven to highly discordant genomic locations during the reprogramming to pluripotency (this study) as well as when forcibly expressed in mouse ESCs (Aksoy et al., 2013a). Yet mutating specific residues in helix 3 of the HMG box of SOX17 leads to factors with a binding profile reminiscent of SOX2 during reprogramming and in pluripotent cells (eSOX17^{FNV}: Figure 6 or SOX17EK: Aksoy et al., 2013a). This suggests that the inability to target pluripotency genes by counteracting heterodimer formation with Oct4 and other POU family proteins on canonical *SoxOct* DNA elements may be a critical step in the natural evolution of SOX17. That is, the mutations that convert SOX17 into a pluripotency reprogramming are likely deleterious for the native function of the protein. By performing a directed evolution screen in mammalian cells, we took advantage of the functional versatility of *Sox* genes and identified several

(G) The immunofluorescence of iPSC colonies at passage 4 in 2i condition generated with eSOX17 or eSOX2 using antibodies for the pluripotency markers NANOG, SOX2, and OCT4 in comparison with iPSC colonies obtained using SOX2 wild-type controls. Scale bar, 100 μm. See also Figure S4.

(H) We propose that DERBY-seq libraries could provide a broad-based tool for lineage reprogramming for any donor-target cell combination.

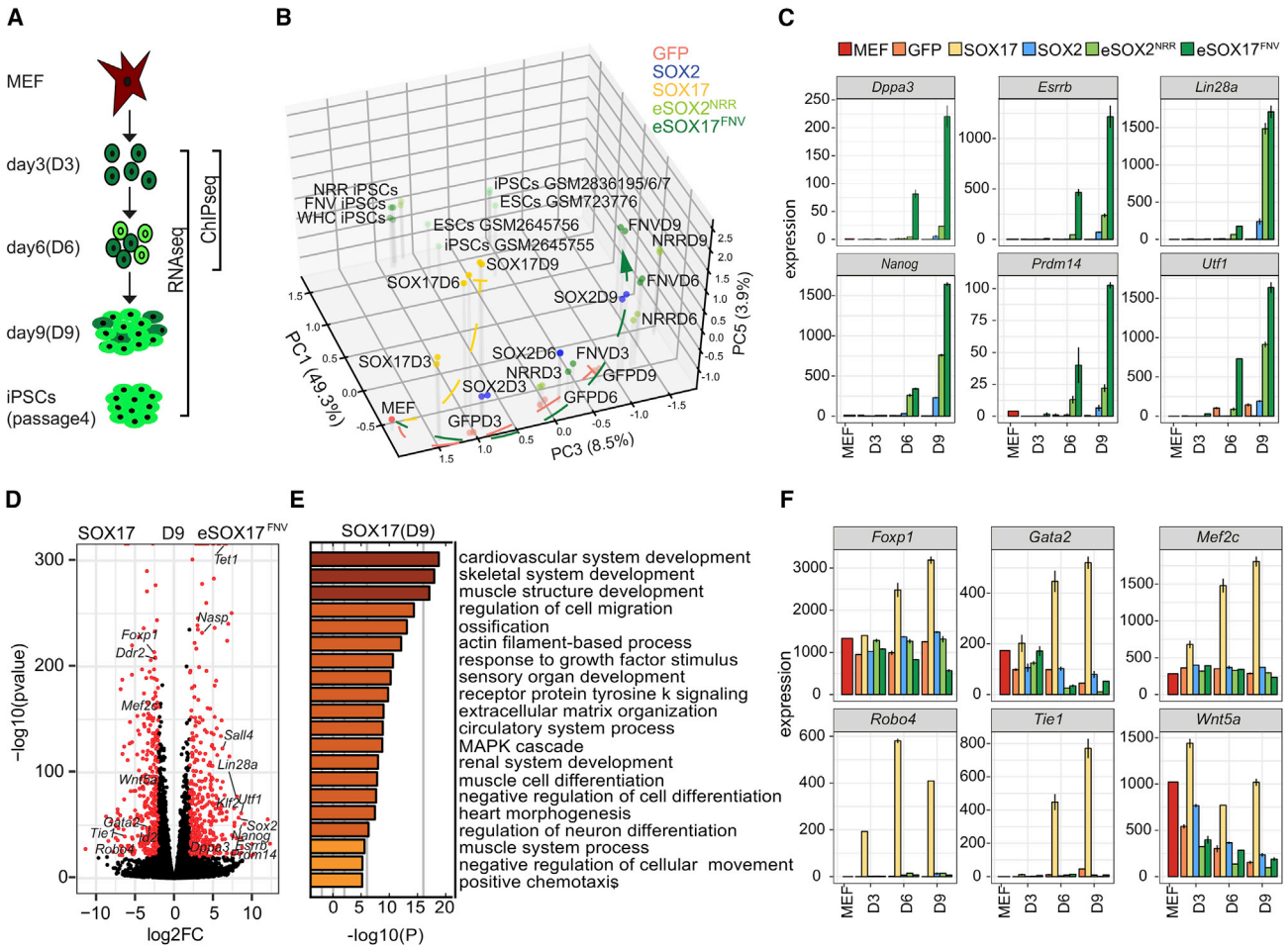


Figure 5. eSOX Variants Accelerate Reprogramming

(A) Experimental flowchart for ChIP-seq and RNA-seq experiments.

(B) Three-principal-component analysis of global gene expression profiles determined by RNA-seq for cells transduced with GFP, SOX2, SOX17, eSOX2^{NRR}, and eSOX17^{FNV} along with OKM at days 3, 6, and 9 was performed using glbase (Hutchins et al., 2014) and PC1, PC3, and PC5. Publicly available MEF, ESC, and iPSC datasets are shown as reference and marked with the GEO identifier (see Table S6). Trajectories are marked with dashed lines. 2D projections are shown in Figure S5A.

(C) The expression levels of selected pluripotency markers are shown as bar plots.

(D) Volcano plot showing differentially expressed genes in eSOX17^{FNV} compared with SOX17 conditions at day 9 of reprogramming. Genes with a log2(fold change) 2 and -log10(p-value) >20 are in red and selected genes are labeled.

(E) Gene ontology analysis for differentially upregulated genes at day 9 in SOX17-OKM-expressing cells compared with eSOX17^{FNV}-OKM performed using metascape (http://metascape.org).

(F) Bar plots for selected genes with a role in the cardiovascular system development that are elevated in the SOX17-OKM condition.

In (C) and (F) the mean of duplicate RNA-seq experiments is shown. The error bars represent the range (maximum to minimum). See also Figures S5 and S6.

artificially evolved and enhanced SOX variants (eSOX) outperforming wild-type SOX2 in three- as well as four-factor cocktails. Our findings demonstrate that native reprogramming factors are not optimally adapted to direct cell-fate conversions that do not occur during embryonic development. For cell-state changes *in vitro* these factors can be profoundly improved by protein design. DERBY-seq is a flexible high-throughput method allowing the robust identification of performance-improving mutations in biomolecule-driven cell-fate conversions. We anticipate that this method will provide a broadly applicable approach for enhancing mammalian cell-fate conversion including the direct lineage reprogramming *in vitro* and *in vivo* (Figure 4H). Upon further refinement of the randomization strategy and selection of appropriate molecular scaffolds, a single multi-purpose DERBY-seq library may be sufficient to identify

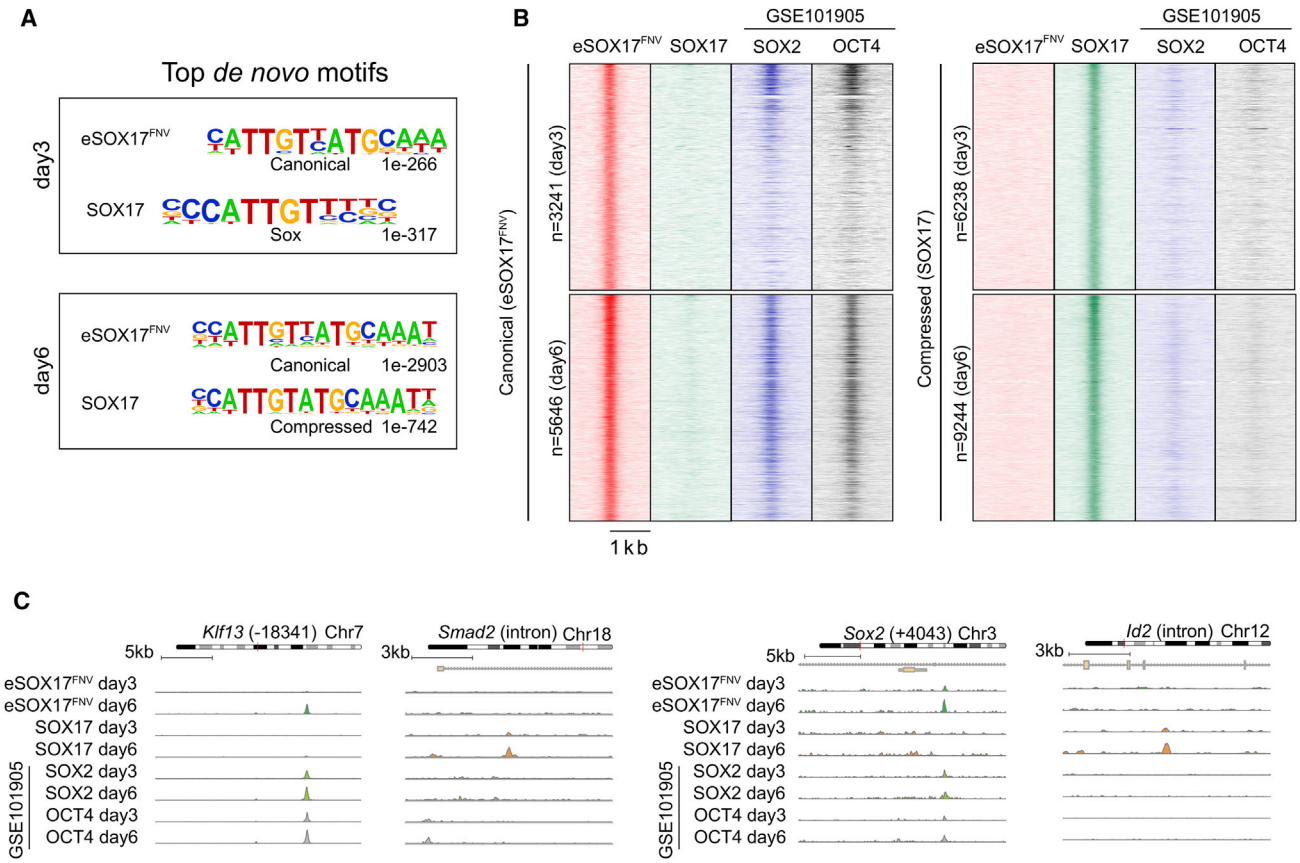


Figure 6. SOX17 and eSOX17^{FNV} Prefer Different DNA Motifs and Target Different Genomic Locations

(A) Sequence logos for the top *de novo* motifs in discovered in eSOX17^{FNV} and SOX17 ChIP-seq peak locations at day 3 and day 6 determined using Homer are shown with p values (http://homer.ucsd.edu/homer/motif/).

(B) Heat maps for eSOX17^{FNV}, SOX17, SOX2, or OCT4 ChIP-seq signals at genomic locations bound by eSOX17^{FNV} containing matches to the canonical *SoxOct* motif (left panel) and by SOX17 containing matches to the compressed *SoxOct* motif (right panel) at day 3 (upper panel) day 6 (lower panel). SOX2 and OCT4 ChIP-seq data are from a study by Knaupp et al. (2017). see also Table S6.

(C) Genome browser plots prepared using gviz (https://bioconductor.org/packages/release/bioc/html/Gviz.html) for selected locations where eSox17^{FNV} is bound to canonical motifs at locations annotated as pluripotency super-enhancers (*Klf13*, *Sox2*; Whyte et al., 2013), or compressed elements bound by SOX17 near *Smad2* (a location previously validated to possess enhancer activity; Aksoy et al., 2013a) or *Id2*.

reprogramming factors for any donor/target cell combination. The only limitation would be the ability to reliably select for cells with the desired phenotype.

EXPERIMENTAL PROCEDURES

Site Selection and Library Construction

On the basis of structural modeling and molecular dynamics simulations (Jauch et al., 2011; Merino et al., 2014; Palasingam et al., 2009), three homologous residues of the third helix of the mouse HMG box of SOX2 and SOX17 involved in the DNA-dependent heterodimerization with POU factors were selected for comprehensive randomization mutagenesis (corresponding to HMG box residues E46/I53/K56 of SOX2 and L46/V53/E56 of SOX17). Each residue was randomized using NNK codons (where N represents A, C, G, or T and K represents G or T) resulting in a library of 20^3 = 8,000 protein variants excluding STOP codon-containing variants (see Figures 1 and S1C). Libraries were generated using the retroviral pMX vector backbone (Kitamura et al., 2003). Randomization and library generation was performed by GENEWIZ (Suzhou, China). One hundred microliters of the bacterial suspension containing pMX plasmids was diluted with Luria broth (LB) medium, plated on 15-cm LB/ampicillin plates, and grown for approximately 18 hr, after which the cells were harvested and cultured in 100 mL of liquid culture LB for another 18 hr. Maxiprep plasmid DNA preparations were performed with the EndoFree Maxi prep kit (Tiangen, #DP117).

Library Transfection and Virus Preparation

Plat-E cells (Morita et al., 2000) were thawed and cultured in Plat-E medium composed of DMEM (Thermo Fisher) supplemented with

10% fetal bovine serum (FBS; Natocor, Argentina) in 10-cm cell culture dishes for at least 36 hr without changing the medium. Cells were passaged every 2–3 days at 70%–80% confluence. Approximately 7–8 million cells per 10-cm plate were seeded 12–16 hr prior to transfection. At 70%–80% confluence, 10 μg of each pMX plasmid was used to transfect Plat-E cells with 40 μg of the transfection reagent polyethyleneimine (Polysciences, #23966) dissolved in 1 mL of Opti-MEM (Thermo Fisher Scientific, #31985070). After 12 hr the medium was changed. Virus-containing medium was collected at 48 and 72 hr after transfection and passed through 0.45-μm filters (Millipore).

Pluripotency Reprogramming

MEFs (OG2-MEF from mouse embryos collected at embryonic day 13.5 carrying a transgenic GFP reporter driven by a *Oct4* promoter; Szabo et al., 2002; Yeom et al., 1996) were obtained from the GIBH animal facility. Animal care and experimental protocols were approved by the Guangzhou Institutes of Biomedicine and Health Ethical Committee. Cells were seeded at ~30,000 cells per well of a 6-well plate or ~15,000 cells per well of a 12-well plate, and cultured in MEF medium composed of high-glucose DMEM containing 4.5 g/L D-glucose supplemented with 10% FBS (Natocor, #SFBE), 1× GlutaMAX (Thermo Fisher Scientific, #35050061), 1× nonessential amino acids (NEAA; Thermo Fisher Scientific, #11140050) and 0.5× penicillin/streptomycin (Hyclone, #SV30010) for 8–10 hr prior to viral transduction. One milliliter of filtered retroviral supernatant containing polybrene at a concentration of 8 mg/mL (Sigma-Aldrich, #40804ES76) of each factor (OSKM) was added twice in a 24-hr interval. After 48 hr of viral infection, MEF medium was replaced with mES medium (high-glucose DMEM containing 4.5 g/L D-glucose supplemented with 15% FBS, 1% NEAA, 1% GlutaMAX, 1% sodium pyruvate, 0.5% penicillin/streptomycin, 1,000 U/mL LIF, 0.055 mM β-mercaptoethanol, and 50 μg/mL vitamin C; Esteban et al., 2010). The day of media change is considered as reprogramming day 0. The reprogramming cells were maintained at 37°C and 5% CO_2 (BB15 incubator; Thermo Fisher Scientific) and monitored using a phase-contrast microscope (Zeiss Axio Vert.A1). Every 24 hr mES medium was changed. For whole-well scanning the medium was removed and 1× Dulbecco's PBS (DPBS; Thermo Fisher Scientific, #14190144) was added (at day 10 for 4F and day 12 for 3F eSOX2 variants, and at day 10 for eSOX17 variants), and whole-well scans were taken from 12-well plates using an ImageXpress Micro XLS confocal High-Content Analysis System (Molecular Devices).

Fluorescence-Activated Cell Sorting

To separate reprogramming from non-reprogramming populations in pooled library screens we performed FACS, whereby OG2-MEFs were reprogrammed in 6-well tissue culture plates for 12–15 days, mES medium was removed, and cells were washed twice with 1× DPBS. The cells in each well were then dissociated with 1 mL of 0.25% trypsin/1 mM EDTA (Thermo Fisher Scientific, #25300054), passed through a 40-μm BD cell strainer, and diluted in FACS buffer (1× DPBS + 2 mM EDTA + 0.1% BSA) to 6–7 million cells/mL. For each sample, cells from three replicate wells were combined in one tube and used for two-way cell sorting. Cells were sorted by using the 488-nm GFP laser channel of a Beckman Coulter-MoFlo Astrios. Approximately, 20,000–100,000 GFP-positive and GFP-negative cells were collected for each sample. To compare eSOX variants, we performed analytical cell sorting using a BD Accuri C6 device with FlowJo 7.6 software analyzing ~30,000 life cells per variant. FACS plots used to calibrate gating are shown in Figures 2B and S1G (lower panel) (BD Accuri) and Figure S2A (MoFlo).

Next-Generation Library Preparation and Sequencing

Genomic DNA was isolated from GFP-positive and GFP-negative cells using a Quick gDNA micro prep kit (Zymo Research, #D3020). As a control, gDNA was also extracted from unsorted, transduced cells 60 hr after transfection (for the eSOX17 library only). As a further control, the Maxiprepped library in the pMX backbone was sequenced for both eSOX libraries. For the eSOX2 input library control, the plasmid library was serially diluted prior to the PCR starting from 5.68^{10} to 5.68^{6} molecules (Figure S2E). For the eSOX17 experiment, the plasmid library was diluted to 1 million molecules per PCR reaction (~0.625 pg).

Amplicon libraries were produced in a three-step (eSOX2 library) or two-step (eSOX17 library) PCR scheme (Figures S2E and S2F). First, pMX transgenes were amplified by a 15-cycle PCR using DreamTaq Green PCR Master Mix (Thermo Fisher Scientific, #K1082) and products were purified using a PCR purification kit (Tiangen, #DP209). Second, a 6-cycle PCR was performed with primers flanking the randomized portion and overhangs encoding barcodes and parts of the adapters required for Illumina sequencing (Table S1). Third, in a last 6-cycle PCR, the remainder of the Illumina adapters was added. The resulting ~250-bp PCR products were electrophoresed and purified using a Midi Gel Purification kit (Tiangen, #DP209). In the case of the eSOX17 library, the exon-intron gene structure allows for the discrimination of endogenous and exogenous *Sox17* and primers flanking the randomized portion and overhangs with adapters were used in the first 15-cycle PCR reaction. In the second 12-cycle PCR, the full barcoded Illumina adapters were added. Samples were quantified with a Qubit (Thermo Fisher Scientific) and 50 ng of DNA was submitted to WuxiApptech for sequencing with an Illumina HiSeq 2500 with cluster generation (concentration 8 pM^2) by using 125-bp paired-end reads. The sample was run using Illumina standard procedures, with 15% genomic PhiX DNA (Illumina) added to increase sequence diversity. Primer sequences used to extract exogenous genes and Illumina adaptors with barcodes used for multiplexing are listed in Table S1.

Colony Picking and Passaging

iPSC colonies with compact dome-shaped morphology were picked between days 10 and 12 using a sterile glass rod and micropipette and transferred into a 1.5-mL tube containing ~30 μL of 0.25% trypsin/EDTA. The cells were then incubated for 3–5 min and seeded on feeder MEFs treated with mitomycin C and grown for 4–5 days in mES medium. Colonies were selected and picked based on dome-shaped morphology and bright GFP fluorescence and seeded on gelatin-coated 24-well plates. From passage 2 onward the cells were maintained in feeder-free conditions in chemically defined 2i medium (Ying et al., 2008) (a 1:1 mix of high-glucose DMEM/F12 [Thermo Fisher Scientific, #C11320500BT]

and Neurobasal medium [Thermo Fisher Scientific, #21103049] containing 1× N2 [Thermo Fisher Scientific, #17502048], 1× B27 [Thermo Fisher Scientific, #17504044], 1× GlutaMAX [Thermo Fisher Scientific], 1× NEAA [Thermo Fisher Scientific], 1 mM sodium pyruvate [Thermo Fisher Scientific], 0.055 mM β-mercaptoethanol [MP Biomedicals], 0.5× penicillin/streptomycin, 1,000 U/mL LIF, 3 μM CHIR99021 [Selleck, #S2924 25 mg], and 1 μM PD0325901 [Selleck, #S1036 25 mg]).

Karyotyping

iPSCs were cultured on 6-cm plates in 2i medium. At 70% confluence, demecolcine (Aladdin, #477305) was added to a concentration of 20 μg/mL. After 1 hr cells were trypsinized, collected by centrifugation at 200 × *g* for 3 min, resuspended in 8 mL of 0.075 M KCl, and incubated at 37°C for 20 min. Two milliliters of fixative solution (acetic acid [Merck Millipore, #100062] and methanol [Merck Millipore, #822283] at 1:3) were added, mixed gently, and incubated at 37°C for 10 min. The supernatant was removed by centrifugation and the pre-cooled fixative solution was added to 10 mL. Cells were distributed on a cold cover slide and incubated at 75°C for 3 hr. After trypsin treatment and Giemsa staining (Sigma-Aldrich, #48900), metaphase spreads were analyzed on a microscope (Olympus BX51).

Immunofluorescence

After 4–5 passages, the iPSCs were counted and 1–2 million cells were seeded on 24-well cell culture dishes pre-coated with 0.1% gelatin, and grown for 24–48 hr in 2i medium until 80% confluence. Cells were washed three times with DPBS (1×) and fixed with 4% paraformaldehyde at room temperature for 15 min. Cells were permeabilized by incubation with 0.2% Triton X-100 (Sigma-Aldrich, #T8787) and dissolved in 10% BSA (MPBIO, #0218054991) in 1× DPBS at room temperature for 30 min. Afterward, the permeabilized cells were washed twice with 1× DPBS and incubated with primary antibodies for NANOG (Novus, #NB100-58842, 1:500), OCT4 (Santa Cruz Biotechnology, #sc-5279, 1:500) and SOX2 (Santa Cruz; #sc-17320, 1:500) at 4°C overnight. The cells were then washed three times for 5 min with DPBS (1×) and incubated with secondary antibodies required for the respective primary antibody (donkey-anti-rabbit: Thermo Fisher Scientific #A24870, 1:250; rabbit-anti-mouse: Thermo Fisher Scientific #A21063, 1:500; donkey-anti-goat: Abcam #ab6949, 1:500) in darkness at room temperature for 2 hr. Cells were washed three times with DPBS (1×) for 5 min. The cells were further stained with 1× DAPI (Thermo Fisher Scientific, #R37606) and imaged with an Axio Vert.A1 (Zeiss).

mRNA Isolation and Real-Time qPCR

The iPSCs after 4–5 passages were cultured in 6-well plates and at 80% confluence the medium was removed, cells were washed with DPBS, and the RNA was extracted using the TRIzol method. Total RNA was purified using the PureLink RNA MiniKit (Ambion, #12183025) and quantified using a Nanodrop spectrophotometer. Five micrograms of total mRNA was used to synthesize the cDNA with a ReverTra Ace qPCR RT master mix (Toybo FSQ-201s). qPCR was performed with iTaq Universal SYBR Green Supermix on a CFX-96 thermocycler (Bio-Rad). Samples were run in technical triplicates. Relative gene expression was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) with *Actin* as endogenous control and the MEFs of 5-day samples as calibrator. Primers are listed in Table S5.

Estimation of Cell Proliferation

OG2 MEF cells were transduced with retroviral supernatant and at reprogramming day 1 cells were trypsinized. The cells from single-cell suspension were counted using a Scepter 2.0 cell counter (Millipore), and 10,000 cells were seeded on 12-well cell culture plates and cultured for 48 hr in mES medium before being counted again.

Processing of Amplicon Sequencing Data

Reads were de-multiplexed using the raw fastq files and a custom Python script. The data were further analyzed using the R (https://www.rstudio.com/), BioStrings (https://bioconductor.org/packages/release/bioc/html/Biostrings.html), and data.table (https://cran.r-project.org/web/packages/data.table/index.html) packages. First, relevant positions of randomized codons were extracted and translated, and trip-peptides were counted to construct count matrices where rows are variants and columns replicate GFP-positive and -negative conditions. DESeq2 (Love et al., 2014) was used to score the differential enrichment of variants in GFP-positive or -negative cell populations by pairing biological replicates (as in Tables S2 and S3). Candidates were selected for validation experiments from volcano plots where $\log_2$ fold change and adjusted p-value scores are taken into account using data from 3F to 4F conditions. Biophysical properties of amino acids were also considered in order to test diverse candidates (i.e., charge, size, hydrophobicity).

ACCESSION NUMBERS

All raw and processed data reported here from DERBY-seq, RNA-seq, and ChIP-seq have been deposited to NCBI GEO under accession number GEO: GSE107987. The Python scripts use to de-multiplex the reads and the R codes used to perform the DERBY-seq analysis are available on request from the authors.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at https://doi.org/10.1016/j.stemcr.2018.07.002.

AUTHOR CONTRIBUTIONS

V.V. and R.J. designed the study, analyzed data, and prepared figures. V.V. and J.O.A. performed experiments and analyzed data. R.J. and V.V. wrote the manuscript. Y.S. provided structural illustrations. V.M. performed ChIP-seq experiments. Y.S., M.W., and X.Y. contributed to reprogramming experiments. All authors read, contributed to, and approved the final manuscript.

ACKNOWLEDGMENTS

R.J. is supported by the Ministry of Science and Technology of China (grant nos. 2013DFE33080 and 2016YFA0100700), the National Natural Science Foundation of China (grant nos. 31471238,

31771454, and 31611130038), a 100 talents award of the Chinese Academy of Sciences and Science and Technology Planning Projects of Guangdong Province, China (2017B030314056 and 2016A050503038). V.V. and V.M. are supported by CAS-TWAS President's Fellowship and University of Chinese Academy of Sciences (UCAS). Y.S. is supported by a Chinese Government Scholarship (CGS) and the University of Chinese Academy of Sciences (UCAS). We thank Guo Wenjing for technical assistance in whole-well scanning.

Received: January 20, 2018
Revised: July 5, 2018
Accepted: July 9, 2018
Published: August 2, 2018

REFERENCES

Aksoy, I., Jauch, R., Chen, J., Dyla, M., Divakar, U., Bogu, G.K., Teo, R., Leng Ng, C.K., Herath, W., Lili, S., et al. (2013a). Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. EMBO J. *32*, 938–953.

Aksoy, I., Jauch, R., Eras, V., Chng, W.B., Chen, J., Divakar, U., Ng, C.K., Kolatkar, P.R., and Stanton, L.W. (2013b). Sox transcription factors require selective interactions with Oct4 and specific transactivation functions to mediate reprogramming. Stem Cells *31*, 2632–2646.

Arnold, F.H. (2015). The nature of chemical innovation: new enzymes by evolution. Q. Rev. Biophys. *48*, 404–410.

Bowles, J., Schepers, G., and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Dev. Biol. *227*, 239–255.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Esteban, M.A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S., et al. (2010). Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. Cell Stem Cell *6*, 71–79.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741–746.

Graf, T., and Enver, T. (2009). Forcing cells to change lineages. Nature *462*, 587–594.

Hou, L., Srivastava, Y., and Jauch, R. (2017). Molecular basis for the genome engagement by Sox proteins. Semin. Cell Dev. Biol. *63*, 2–12.

Hutchins, A.P., Jauch, R., Dyla, M., and Miranda-Saavedra, D. (2014). glbase: a framework for combining, analyzing and displaying heterogeneous genomic and high-throughput sequencing data. Cell Regen.(Lond.) *3*, 1.

Jauch, R., Aksoy, I., Hutchins, A.P., Ng, C.K., Tian, X.F., Chen, J., Palasingam, P., Robson, P., Stanton, L.W., and Kolatkar, P.R. (2011). Conversion of Sox17 into a pluripotency reprogramming factor by reengineering its association with Oct4 on DNA. Stem Cells *29*, 940–951.

Jerabek, S., Ng, C.K., Wu, G., Arauzo-Bravo, M.J., Kim, K.P., Esch, D., Malik, V., Chen, Y., Velychko, S., MacCarthy, C.M., et al. (2017). Changing POU dimerization preferences converts Oct6 into a pluripotency inducer. EMBO Rep. *18*, 319–333.

Kitamura, T., Koshino, Y., Shibata, F., Oki, T., Nakajima, H., Nosaka, T., and Kumagai, H. (2003). Retrovirus-mediated gene transfer and expression cloning: powerful tools in functional genomics. Exp. Hematol. *31*, 1007–1014.

Knaupp, A.S., Buckberry, S., Pflueger, J., Lim, S.M., Ford, E., Larcombe, M.R., Rossello, F.J., de Mendoza, A., Alaei, S., Firas, J., et al. (2017). Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. Cell Stem Cell *21*, 834–845.e6.

Liu, Y., Asakura, M., Inoue, H., Nakamura, T., Sano, M., Niu, Z., Chen, M., Schwartz, R.J., and Schneider, M.D. (2007). Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells. Proc. Natl. Acad. Sci. USA *104*, 3859–3864.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods *25*, 402–408.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Malik, V., Zimmer, D., and Jauch, R. (2018). Diversity among POU transcription factors in chromatin recognition and cell fate reprogramming. Cell Mol Life Sci *75*, 1587–1612.

Merino, F., Ng, C.K.L., Veerapandian, V., Scholer, H.R., Jauch, R., and Cojocaru, V. (2014). Structural basis for the SOX-dependent genomic redistribution of OCT4 in stem cell differentiation. Structure *22*, 1274–1286.

Morita, S., Kojima, T., and Kitamura, T. (2000). Plat-E: an efficient and stable system for transient packaging of retroviruses. Gene Ther. *7*, 1063–1066.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26*, 101–106.

Ng, C.K., Li, N.X., Chee, S., Prabhakar, S., Kolatkar, P.R., and Jauch, R. (2012). Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. Nucleic Acids Res. *40*, 4933–4941.

Packer, M.S., and Liu, D.R. (2015). Methods for the directed evolution of proteins. Nat. Rev. Genet. *16*, 379–394.

Palasingam, P., Jauch, R., Ng, C.K., and Kolatkar, P.R. (2009). The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. J. Mol. Biol. *388*, 619–630.

Remenyi, A., Tomilin, A., Pohl, E., Lins, K., Philippsen, A., Reinbold, R., Scholer, H.R., and Wilmanns, M. (2001). Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. Mol. Cell *8*, 569–580.

Schuetz, A., Nana, D., Rose, C., Zocher, G., Milanovic, M., Koenigsmann, J., Blasig, R., Heinemann, U., and Carstanjen, D. (2011). The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. Cell Mol Life Sci *68*, 3121–3131.

Szabo, P.E., Hubner, K., Scholer, H., and Mann, J.R. (2002). Allele-specific expression of imprinted genes in mouse migratory primordial germ cells. Mech. Dev. *115*, 157–160.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Tanabe, K., Haag, D., and Wernig, M. (2015). Direct somatic lineage conversion. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140368.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Yeom, Y.I., Fuhrmann, G., Ovitt, C.E., Brehm, A., Ohbo, K., Gross, M., Hubner, K., and Scholer, H.R. (1996). Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. Development *122*, 881–894.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., et al. (2015). A XEN-like state bridges somatic cells to pluripotency during chemical reprogramming. Cell *163*, 1678–1691.