832–845 *Nucleic Acids Research, 2017, Vol. 45, No. 2* *Published online 3 December 2016*
doi: 10.1093/nar/gkw1198

Quantitative profiling of selective Sox/POU pairing on hundreds of sequences in parallel by Coop-seq

Yiming K. Chang[1], Yogesh Srivastava[2,3,4], Caizhen Hu[2,3,4], Adam Joyce[5], Xiaoxiao Yang[2,3,4], Zheng Zuo[1], James J. Havranek[5], Gary D. Stormo[1,*] and Ralf Jauch[2,3,4,*]

[1]Department of Genetics and Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA, [2]Genome Regulation Laboratory, Drug Discovery Pipeline, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China, [3]Key Laboratory of Regenerative Biology, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China, [4]Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China and [5]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA

Received October 17, 2016; Revised November 14, 2016; Editorial Decision November 17, 2016; Accepted November 17, 2016

ABSTRACT

Cooperative binding of transcription factors is known to be important in the regulation of gene expression programs conferring cellular identities. However, current methods to measure cooperativity parameters have been laborious and therefore limited to studying only a few sequence variants at a time. We developed Coop-seq (cooperativity by sequencing) that is capable of efficiently and accurately determining the cooperativity parameters for hundreds of different DNA sequences in a single experiment. We apply Coop-seq to 12 dimer pairs from the Sox and POU families of transcription factors using 324 unique sequences with changed half-site orientation, altered spacing and discrete randomization within the binding elements. The study reveals specific dimerization profiles of different Sox factors with Oct4. By contrast, Oct4 and the three neural class III POU factors Brn2, Brn4 and Oct6 assemble with Sox2 in a surprisingly indistinguishable manner. Two novel half-site configurations can support functional Sox/Oct dimerization in addition to known composite motifs. Moreover, Coop-seq uncovers a nucleotide switch within the POU half-site when spacing is altered, which is mirrored in genomic loci bound by Sox2/Oct4 complexes.

INTRODUCTION

The Sry-related box (Sox) and Pit-Oct-Unc (POU) families of transcription factors (TFs) are critical regulators of gene expressions programs that determine cellular identities. The mouse and human genomes encode 20 Sox and 14 POU genes (1,2). Sox factors possess a 79 amino acid high-mobility group (HMG) box allowing them to bind the minor groove of the DNA in a sequence specific manner with CATTGTC-like consensus sequences (3–6). Most POU factors predominantly bind to an octamer ATGCTAAT-like consensus sequence leading to their designation as Oct proteins. DNA binding is accomplished with a bi-partite DNA binding domain (DBD) consisting of a N-terminal POU-specific (POU_S) binding the ATGC and a C-terminal POU-homeodomain (POU_{HD}) binding the TAAT part of the octamer element (7,8). The DBDs of Sox and POU TFs not only mediate DNA recognition but also DNA-dependent heterodimerization between members of the two protein families (reviewed in (9)). The Sox2/Oct4 heterodimer has been most extensively studied as it represents the core component of a gene regulatory network in pluripotent stem cells. The Sox2/Oct4 complex co-binds 1000's of enhancers in mouse and human embryonic stem cells using a composite DNA element with directly juxtaposed half-sites (CATTGTC/ATGCTAAT) referred to as the 'canonical' Sox/Oct motif (10–14). Yet, a subset of Sox2/Oct4 target genes, such as *Fgf4,* is regulated via a variant Sox/Oct element with 3 bp spacer (CATTGTCnnnATGCTAAT) (15). The Sox2/Oct4 partnership is not only crucial for maintaining but also for inducing pluripotency during cellular reprogramming of somatic cells to induced pluripotent stem

*To whom correspondence should be addressed: Tel: +86 2032093805; Fax: +86 2032093805; Email: ralf@gibh.ac.cn
Correspondence may also be addressed to Gary D. Stormo. Tel: +1 314 747 5534; Email: stormo@wustl.edu

cells (iPSCs) (16–19). As individual Sox and POU factors bind near-identical DNA sequences but target and regulate unique sets of genes, the selective partnerships of these factors has been suggested to determine their specific developmental functions by means of a 'partner code' (20,21). Indeed, a Sox17/Oct4 heterodimer was found to bind to a variant 'compressed' CATTGT/ATGCAAAT sequence lacking the terminal base-pair (bp) of the Sox half-site to direct the differentiation of the primitive endoderm (14). Moreover, the Sox17/Oct4 heterodimer is thought to direct the specification of the human germ cell lineage potentially employing similar sequence signatures (22). Group III POU factors including Oct6, Brn2 and Brn4 are co-expressed with Sox2 in a variety of neural lineages (Supplementary Figure S1). Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) studies showed that Sox2 pairs with Brn2 on sequences resembling the 'canonical' motif during the differentiation of neural progenitor cells (23). Additional Sox factors including Sox5, Sox15, Sox17 and Sox18 are also frequently co-expressed with POUIII TFs in neural lineages (Supplementary Figure S1) although their molecular interplay has not yet been established. Importantly, mutations that influence Sox/Oct heterodimerisation can profoundly change the cell fate choices directed by these proteins. For example, we have previously demonstrated that the structure-based engineering of the interaction of Sox2 and Sox17 with Oct4 by 1–2 point mutations swaps how these TFs partner with Oct4 to direct cell fate decisions (24–26). A point-mutated Sox17E57K cooperatively dimerizes with Oct4 on the canonical rather than the compressed sequence and acquires the activity of wild-type Sox2 to induce pluripotency in somatic cells (24). To enable Sox2 to effectively dimerize with Oct4 on the compressed element the reciprocal Sox2K57E mutation is not sufficient. Rather, a Sox2E46LK57E double mutation is required and installs Sox17-like heterodimerization with Oct4 on the compressed element (25). Whether the engineered Sox2E46LK57E factor can replace for Sox17 in a cellular environment has not yet been tested. Using iPSC generation and pluripotency maintenance assays, further Sox2 mutations have been studied (27). For example, a Sox2R75E that modulates Sox2/Oct4 heterodimerization on *Fgf4*-like but not canonical DNA elements (28) was found to reduce, but not abrogate, the potency of Sox2 to maintain pluripotency (27). Clearly, cooperative associations of these factors crucially influences which genomic loci they bind, which sets of genes they regulate and which cellular fate decisions are made in response to this activity. In light of the frequent co-expression of Sox/Oct pairs, the presently known examples of pairing Sox/Oct factors and the known composite DNA elements are likely an incomplete list.

Biophysical cooperativity can be due to direct protein-protein interactions or can be communicated allosterically via the DNA (29). This has to be distinguished from the indirect cooperativity mediated by nucleosomes (30). Direct cooperativity provides an increase in occupancy of specific DNA segments compared to independent binding because the actual binding affinity of a TF for a particular site is increased in the presence of the cooperating TF. The affinity can be increased many-fold, allowing for even intrinsically weak binding sites to be highly occupied when both TFs are available and it facilitates regulatory switches where the occupancy is sharply altered by the TF concentration. Inferences of co-regulating TFs are generally based on co-occurrences of their binding sites within short stretches of DNA sequence (31,32). If the spacing and the orientation of the monomeric binding sites is constrained, direct cooperativity is assumed (33). The importance of composite DNA sequences targeted by TF complexes has led to the adaptation of techniques such as protein-binding microarrays (34) and SELEX-seq (35). More recently, the scale of these investigations has been greatly increased through CAP-SELEX, which enabled screening of large collections of TFs by identifying over-representation of specific orientation and spacing of the independent binding sites (36). These methods are invaluable to identify novel co-motifs and showed that composite binding elements are more than linear combinations of the individual consensus motifs of participating partners. Rather, TF dimerization can profoundly modify the binding specificity of the TFs involved. While the CAP-SELEX method increases the throughput to discover TF pairs with apparent cooperative binding, it does not provide accurate quantitation of the cooperativity parameters or how they vary depending on the sequence. The SELEX-seq approach of Slattery *et al.* (35) can provide good estimates of those parameters by fitting a biophysical model to the occurrences of different sequences after multiple rounds of selection for sites bound to both proteins. A microfluidic approach has also been adapted to the study of cooperatively binding TFs (37). That method can determine both affinity and cooperativity parameters, but requires sophisticated equipment and some complex modeling because not all states of the DNA are measured simultaneously. Here, we describe Coop-seq, which allows for the efficient and accurate determination of relative affinity and cooperativity to a large collection of sequences in parallel and requiring only standard lab equipment and access to DNA sequencing. No modeling is required because it directly measures (to the accuracy allowed by EMSA) the relative occupancy of each possible state of the DNA, unbound, singly bound by each protein, and doubly bound by both proteins. While EMSA has been used previously to measure cooperativity (26,38–39) Coop-seq increases the throughput by over 300-fold. We use Coop-seq to analyze the interactions between 12 pairs of the Sox- and POU-family proteins and demonstrate that Coop-seq reveals highly accurate cooperativity parameters. Using the Coop-seq methodology we uncover a series of novel features facilitating DNA recognition by Sox and POU TF families.

MATERIALS AND METHODS

Protein expression and purification

Sox2, protocol 1. The 79 amino acids of the HMG box of the mouse Sox2 proteins (accession BC057574.1) were cloned into a pUC19 based plasmid with T7 promoter and T7 terminator containing N-terminal strep-tag followed by cleavage site for thrombin protease as described (40). The construct was transformed into *Escherichia coli* BL21(DE3) and grown in Luria broth (LB). Protein expression was induced by adding 0.4 mM isopropyl-B-thiogalactoside (IPTG) for 3 h at 30°C. The proteins were purified using

Strep-Tactin Superflow (IBA Life Sciences) following the manufacturer's instructions. The strep-tag was cleaved off by thrombin protease digestion for 8 h at room temperature.

Oct4, protocol 1. DNA sequence of the POU domain of mouse Oct4 (accession NM_001252452.1; residues 1–156), codon-optimized for expression in *E. coli*, was amplified by PCR and cloned into the pET-42(a+) vector containing a C-terminal thrombin cleavage site followed by an 8x-His affinity tag. The construct was transformed into *E. coli* Arctic Express (DE3) cells (Agilent), which were grown at 37°C in Terrific Broth (TB) supplemented with 50 mM glucose and 50 g/ml Kanamycin until OD (600 nm) reached 0.4. Washed cells were transferred to fresh 1× TB with no added glucose for 3 h of growth at 25°C. Protein expression was carried out for 8 h following the addition of 0.4 mM IPTG.

Inclusion bodies were extracted by the method of Palmer and Wingfield (41). Harvested cells were sonicated in ethylenediaminetetraacetic acid (EDTA)-free wash buffer, and the insoluble material was resuspended in wash buffer (WB) [100 mM Tris 7.5; 10 mM β-mercaptoethanol (b-ME), 2 M urea, 2% Triton-X] three times. For the final wash, both urea and b-ME were omitted, and the precipitate was resuspended and incubated in Extraction Buffer [50 mM Tris–HCl pH 7.5; 8 M guanidine–HCl]. Supernatant was applied to Ni-NTA agarose (Qiagen) and eluted using a pH gradient. Eluent was dialyzed into refolding buffer (RB) [Tris–HCl pH 7.5; 150 mM NaCl; 10% glycerol] supplemented with 2 M guanidine–HCl, and protein was re-folded by rapid dilution at 4°C in stirred RB. The refolding mixture was concentrated by dialysis against a 20% PEG-20k gradient, and the affinity tag was removed by thrombin protease. Cleavage product was further purified by gel filtration chromatography (Sephacryl S-100 HR, GE Healthcare Life Sciences) in RB to remove thrombin protease the cleaved affinity tag and trace denaturant. Fractions with protein of interested were pooled and concentrated by centrifugation (Amicon, MWCO 10k).

Sox, POU protocol 2. HMG boxes of Sox2, Sox5, Sox15, Sox17, Sox18 and mutants and the POU domain of OCT4 were also produced using an alternative protocol with pDEST-hisMBP (42) or pETG20A vectors and the *E. coli* expression systems as described (6,24,26). The POU domains of Brn2 and Oct6 were produced as reported in (43). Following these established protocols, the POU domain of Brn4 was newly prepared for the present study. In brief, the mouse Brn4 POU (accession BC138657.1; residues 186–337) was introduced into the GATEWAY destination vector pDEST-hisMBP(41) using the LR reaction (LifeTechnologies). The fusion protein containing an N-terminal His6-MBP tag was expressed in *E. coli* BL21(DE3) cells grown in 1× TB medium supplemented with 0.2% glucose and 100 μg/ml ampicillin at 37°C to OD600 of ∼0.6–0.8. Next, 0.5 mM IPTG was added and the temperature was adjusted to 18°C to allow protein expression for 18–22 h. The cells were harvested by centrifugation and the pellet was resuspended in lysis buffer [20 mM HEPES pH 7.0; 200 mM NaCl; 10 mM b-ME; 1 mM EDTA; 20 mM Imidazole] and disrupted by sonication. The His6MBP-fusion proteins were first captured using Ni-Agarose beads and subsequently cleaved using tobacco etch virus (TEV) protease at 4°C overnight to remove the His6MBP tag. Cleaved protein was further purified using a 6 ml Resource-S (GE Healthcare) column equilibrated in buffer A [20 mM HEPES pH7.0; 100 mM NaCl] connected to an AktaExpress system and eluted with a linear NaCl gradient using Buffer B [20 mM HEPES pH7.0; 1 M NaCl]. Fractions containing pure BRN4-POU protein were pooled, exchanged into a storage buffer [10 mM HEPES pH7.0; 100 mM NaCl] using PD-10 desalting column (GE Healthcare), frozen using liquid nitrogen and stored in aliquots at −80°C.

Library design and preparation

DNA libraries were designed by flanking the degenerate sequences of interest (those in Figure 1C) with 5′ flanking sequence of GAGTCGTCTCGTCAGCAC and 3′ flanking sequence of CCGTAGAGCACTCAGGTC for downstream processing. Libraries were procured by ordering single stranded DNA oligos from IDT. To make double-stranded DNA (dsDNA) libraries, 100 pmol single-strand degenerate template sequences were mixed with an equal amount of reverse complement primer (GACCTGAGTGCTCTACGG). In the presence of Taq Polymerase (Lambda Biotech), brief 10-s denaturing followed by 10 min of 55°C annealing/extension is sufficient to make dsDNA libraries. Because any unextended single-stranded DNA (ssDNA) could contaminate the unbound band, the reaction mix was digested by 1 ml NEB Exo I exo-nuclease (New England Biolabs) for 30 min. All final dsDNA products were purified by PCR purification columns (QIAGEN) and eluted in MilliQ water (Millipore).

Coop-seq experiments

All binding reactions were done in a 10 μl reaction volume using 50 nM Sox and 100 nM Oct proteins, 1 μM of dsDNA library in 1× NEB Cutsmart buffer [50 mM Potassium Acetate; 20 mM Tris-acetate; 10 mM magnesium acetate; 100 μg/ml BSA, pH 7.9 at 25°C] supplemented with 10% glycerol and were incubated for 30 min on ice. Electrophoresis mobility assay (EMSA) were done using native 12% PAGE prepared as Tris/Glycine [25 mM Tris pH 8.3; 192 mM glycine] mini-gels (Bio-Rad). These gels were first pre-run using 1× Tris/glycine buffer at 200 V for 30 min, then samples were loaded and gels were run for an additional 60 min at 200 V at 4°C. After EMSA, the gels were stained with ethidium bromide and visualized using Tanon 1600 or Bio-Rad gel imager. Each band detected in the EMSA were excised with a disposable sterile toothpick and the DNA in the gel extracted by incubating for 30 min at 50°C in 50 μl acrylamide gel extraction buffer [500 mM ammonium acetate; 10 mM magnesium Acetate; 1 mM EDTA; 0.1% sodium dodecyl sulfate (SDS)]. 0.01 pmol of PCR Control DNA with the sequence of GAGTCGTCTCGTCAGCACCGGCGGCGGTTCCCGGAAAGACCGTAGAGCACTCAGGTC (primer binding sites underlined) were added to each extracted sample. Sample in the extraction buffer were purified with QIAquick Nucleotide Removal Kit (Qiagen) following the manufacturer's instructions and recovered using double distilled H_2O (ddH_2O). Each fraction of DNA was

barcoded and amplified using HotStart PCR Master Mix (Lambda Biotech). DNA was denatured at 94°C for 30 s, annealed at 55°C for 30 s and extend at 72°C for 45 s per round for 12–20 rounds with modified Indexed-Illumina primers (PE1-Genetics1/2, PE2.0) (Supplementary Table S3, barcodes shown in boldface). The PCR product was then purified again using QIAquick Nucleotide Removal Kit.

Overall, 50 ng DNA per sample at 4 ng/μl was subjected for sequencing using the Illumina Hiseq2000. Fifteen percent of PhiX genomic DNA was added to each sequencing reaction to increase library complexity. Each Illumina lane provided more than 100 million reads, which ensured sufficient sequencing depth. 125 bp paired end reads were obtained and quality filtered. Libraries were de-multiplexed with custom python scripts. Sequencing results were parsed by counting how many times each variant of oligo were sequenced for each of the excised bands. Only species with read counts of at least five were considered for downstream analysis. The relative cooperativity value (ω_i) is calculated from the reads for sequence $S_i$in each fraction (band of the EMSA gel) as (see Results and Figure 1).

$$\omega_i = \frac{\#S_i\,(f_{dimer})\,\#S_i\,(f_{unbound})}{\#S_i\,(f_{Sox\_})\,\#S_i\,(f_{Oct})}$$

Analysis and visualization of Coop-seq data

Raw read counts were imported into R (https://www.r-project.org/) and transformed using base functions and data.table (http://datatable.r-forge.r-project.org/), tidyr and stringr packages. If not indicated otherwise plots were prepared with the ggplot2 package. For Pearson correlation analysis (Figure 2A) and principle component analysis (PCA, Figure 2C and D) a data matrix was prepared with mean normalized cooperativity values for all 324 sequences as rows and 12 Sox/Oct pairs as columns. Rows with missing values (NA) were omitted resulting in 274 sequences. PCA was performed using the prcomp function in R with center and scale set to TRUE. The pheatmap package and ward.D2 algorithm was used for the hierarchical clustering of the correlation coefficients and the visualization of the correlation heatmap. The heatmap in Figure 4A was prepared using a using a matrix with the 12 Sox/Oct pairs as rows and 23 half-site configurations as columns and the pheatmap package. Average ω values including all replicates and sequence variants were natural log transformed and rows and columns were hierarchically clustered using the ward.D2 algorithm.

Counting motif occurrences in ChIP-seq datasets

Bed files were imported into R using fread (data.table), converted to GenomicRanges objects and sequences were retrieved from mm9, mm10 or rn5 genome versions using BSGenome objects and the getSeq function (Biostrings). IUPAC strings were used to search the ChIPseq-enriched regions in forward and reverse direction using the vcountPattern function (Biostrings) allowing for one mismatch and fixed = F.

Structural modeling

Sox2/Oct4/DNA complexes of all four orientations (FF, FR, RF and RR) were prepared by using a previously generated Sox2/Oct4 model on the canonical Sox/Oct DNA element (GGCATTGTCATGCAAATCGGCGG) as a template. The ternary complexes of alternative Sox2/Oct4/DNA configurations were generated using chimera 1.10.1 (http://www.cgl.ucsf.edu/chimera/) by isolating the Sox2/DNA and Oct4/DNA complexes and superimposing 3 bp DNA of these binary complexes to corresponding central 3 bp DNA fragments of modeled ideal B-DNA (FF:CATTGTTATGCTAAT, FR:CATTGTTATTAGCAT, RF:AACAATGATGCTAAT, RR:AACAATGATTAGCAT; base-pairs used for superposition are underlined). Next, sequences of the binary Sox or Oct-bound DNA elements were converted into corresponding nucleotides in the ideal B-DNA sequence and overlapping nucleotides and the B-DNA template were deleted. The phosphodiester bonds between adjacent bp of Sox2/DNA and Oct4/DNA complexes were created using the Chimera 'adjust bond' option. The energy of the resulting ternary complex models was minimized by using amber force fields ff14SB for the proteins and Bsc0 for DNA. During the minimization process clash scores were counted by using the 'find clashes' option with the default parameter (van der waals (vdw) overla$P = 0.6$ Å, and allowance values 0.4 Å for potential H-bonding pairs). The Oct6 model was generated by threading using the SwissModel server (44) and PDB ID 1GT0 as a template. All structural cartoons were visualised using Chimera.

RESULTS AND DISCUSSION

Outline of the Coop-seq methodology

To probe whether individual Sox and POU proteins exhibit unique preferences for DNA motifs, we first determined the binding specificity of each protein using Spec-seq (45). Sox2 and Sox17 Spec-seq experiments were performed using a DNA library consisting of binding sites CATNNNN and NNNNGTT. Both Sox proteins show similar motifs with a preferred sequence of CATTGTT (Supplementary Figure S2 and Supplementary Table S1). Spec-seq for Brn2, Brn4, Oct6 and Oct4 (also known as Pou3f2, Pou3f4, Pou3f1 and Pou5f1, respectively) was performed using a DNA library with binding sites ATGCNNNN, ATNNNNAT and NNNNTAAT. All four POU proteins show similar motifs with a preferred binding site of ATGC(A/T)AAT (Supplementary Figure S2 and Supplementary Table S2). Given the preferred binding sites of each protein, we can measure cooperativity between protein pairs using Coop-seq which is performed by adding both proteins to a library of DNA sequences, letting the binding reaction come to equilibrium and then separating the microstates (unbound DNA, Sox monomer bound DNA, Oct monomer bound DNA and Sox/Oct dimer bound DNA) on an EMSA gel (Figure 1A). Next, the DNA from each of the bands is extracted and subjected to low-cycle number PCR for amplification, addition of sequencing adaptors and barcodes (Figure 1A). Next, deep sequencing is carried out. For each sequence, $\mathbf{S_i}$, the cooperativity $\boldsymbol{\omega_i}$, is defined as the ratio of the associa-

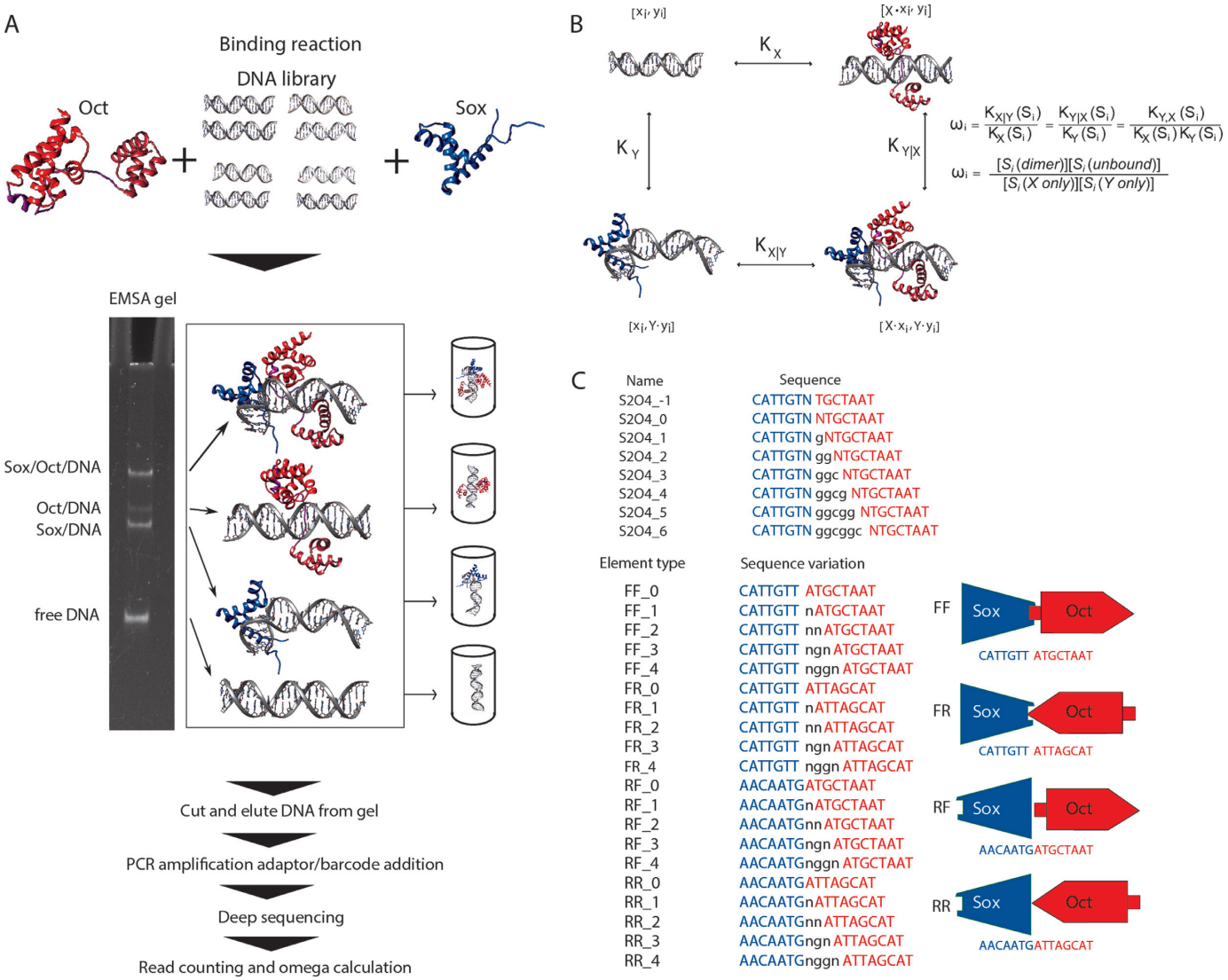

Figure 1. Coop-seq workflow to decipher the Sox/Oct partner code. (**A**) Work flow of Coop-seq. After a binding reaction containing the library of DNA sequences and the DNA-binding proteins attains equilibrium it is run on a EMSA gel to separate the heterodimer bound, monomer bound and unbound fractions. An example lane of an EMSA gel lane stained with ethidium bromide and visualized on using a ChemiDoc XRS+ (Bio-Rad) is shown. Each of the fractions is PCR amplified, barcoded and sequenced through Illumina deep sequencing. (**B**) A general reaction diagram for measuring cooperativity of any two proteins, X and Y to a single sequence S_i with binding sites for each protein x_i, and y_i. Several equivalent equations for cooperativity (ω_i) are shown, including the measurement from the concentrations of the sequence in each band (26,45). K_X denotes the association constant for the binding of protein X alone to sequence S_i. $K_{X|Y}$ denotes the association constant for protein X when protein Y is already bound. (**C**) The binding sites for all sequences used in these Coop-seq experiments. The first library contains sequences with −1 to 6 nucleotide spacers for the Sox and Pou proteins both in the forward direction. Each sequence contains one randomized nucleotide in both Sox2 (blue) and Oct4 (red) binding sites at the position immediately adjacent to the spacer (black). The second library contains all four possible combinations of binding site orientations. Each orientation has 0–4 spacer bases, the outer most of which are randomized. Data from both libraries were combined for the subsequent analysis.

tion constants between the protein binding in the presence of the other protein to binding independently, which can be determined by the ratio of counts for that sequence in the different bands (Figure 1A and B) (26,45)

$$\omega_i \equiv \frac{[S_i\,(dimer)][S_i\,(unbound)]}{[S_i\,(Sox\ only)][S_i\,(Oct\ only)]} \propto \frac{\#S_i\,(f_{dimer})\,\#S_i\,(f_{unbound})}{\#S_i\,(f_{Sox-})\,\#S_i\,(f_{Oct})}$$

where $\#S_i\,(f)$ is the number of reads for sequence $\mathbf{S_i}$ in the indicated fraction. This equation provides the relative cooperativity for each sequence from the read counts in each fraction, but the absolute cooperativity parameters can be obtained in either of two ways. First, one can determine the proportion of the total DNA in each fraction, for example by spiking in a constant amount of a control oligo to each fraction after extraction from the gel, and using that to normalize the reads in each fraction (45). Alternatively, one can use prior knowledge of one of the cooperativity parameters and normalize all other values to this cooperativity value. In this study, we use the forward–forward (FF)+6 configuration previously shown be bound with an $\omega = 1$ by Sox2 and Oct4 (26). Using that normalization gives us absolute cooperativity values for several other specific sequences that are consistent with previously measured values (26). We probed

the cooperativity for the four possible orientations of Sox and Oct half-sites and varied the spacing between them (Figure 1C). The first spacer library includes eight half-site spacing's with spacer lengths of −1 to 6 (with 0 defined as the two binding sites being adjacent as in the 'canonical' Sox/Oct site). The nucleotide in each binding site closest to the spacer is randomized, because previous work has shown that cooperative binding may induce variations in the preferred sequence at the edges of the binding sites (34,35), and the spacers are filled with GGC repeats. The second library contains the four possible arrangements of binding sites: forward/forward (CATTGTC/ATGCTAAT, FF), forward/reverse (FR), reverse/forward (RF), and reverse/reverse (RR). In addition, each orientation contains spacer lengths from 0 to 4 with the nucleotides in the spacer immediately adjacent to the core binding sites being randomized in Sox or Oct half-sites. In total we interrogated the cooperative binding of 12 Sox/Oct pairs to 324 sequences involving 4 POU domains, 9 Sox HMG boxes including four mutant Sox factors leading to 11988 data points including replicate experiments. For the set of sequences and dimers previously analyzed by conventional EMSA (26), the values determined in this work are consistent demonstrating that Coop-seq does not reduce accuracy while increasing the throughput by over 300-fold.

Sox/Oct pairs cluster into distinctive co-binding groups

Cooperativity factors determined by Coop-seq vary over ∼3–4 orders of magnitude for the 12 studied Sox/Oct pairs (Figure 2A). The majority of ω values is <1 (anti-cooperative or negative cooperativity) for all pairs except for the Sox15/Oct4 combination. Curiously, the Sox15/Oct4 dataset is the only pair with a median ω>1 suggesting that it can accommodate a large number of sequences in a moderately cooperative fashion ($1 < \omega < 10$). The majority of Sox/Oct pairs bind with strongly positive cooperativity ($\omega > 10$) to only a small subset of sequences. The sole exception is the mutant Sox2K57E/Oct4 pair which does not bind to any sequence with a ω>10. Next, we generated a Pearson correlation matrix showing that the cooperativities for all 12 Sox/Oct pairs across all binding site variants are positively correlated, but to quite different extents (Figure 2B). Principle component analysis (PCA) was performed as a further technique to identify clusters of dimer-pairs with similar heterodimerization preferences (Figure 2C). These analyses revealed the partitioning of the Sox/Oct pairs into two major groups. The largest one with seven pairs includes all four Sox2 dimers (with Oct4, Brn2, Brn4 and Oct6), Sox15/Oct4, the Sox2R75E and the Sox17E57K mutants. The second cluster comprises Oct4 dimers with the SoxF factors Sox17 and Sox18 as well as the Sox2 double mutant Sox2E46LK57E. Sox5/Oct4 and Sox2K57E/Oct4 could not be unambiguously assigned to either of the two clusters. Thus, Coop-seq confirms our previous observation from classical EMSAs that the dimerization preferences of Sox2 and Sox17 can be swapped with 1–2 point mutations (24–26). Intriguingly, only a small subset of sequences explains most of the variance in the dataset (Figure 2D). A biplot of rotated data projected along PC1 and PC2 reveals that most sequences cluster around the point-of-origin indicating negligible contributions to the differential Sox/Oct dimerization. By contrast, sequences in the FF−1, FF0 and to a lesser extend FF+3 and FF+1 configurations form separate groups indicating selectivity in the recruitment of Sox/Oct pairs. Together, in contrast to Spec-seq that revealed very similar Sox and Oct motifs, Coop-seq identifies distinctive clusters of Sox/Oct dimers. Moreover, only a rather small number of signature sequences are responsible for this variability.

Half-site configurations promoting Sox/Oct partnerships

The distribution of ω values for all 12 Sox/Oct pairs as a function of half-site orientation shows that the FF configuration comprises the broadest range with the strongest deviation from independent binding ($\omega = 1$) (Figure 3A). The majority of measurements for the FR and RR configuration indicated strongly competitive binding (ω<1) whilst the RF configuration included a higher density of positively cooperative binding events ($\omega > 1$). As exemplified for the Sox2/Oct4 dimers (Figure 3B), the cooperativity for the FF configuration is highly dependent on the half site spacing. Sox2/Oct4 exhibits positive cooperativity in the well-studied 'canonical' FF0 configurations whilst binding in the FF−1, FF+1 and FF+2 is on average competitive. In the FF+3 configuration, resembling the classical Sox2/Oct4 binding site from the *Fgf4* promoter (15), binding is moderately cooperative. When half-site orientation is changed, however, cooperative complex formation is severely impeded for FR and RR. This effect can be partially alleviated with increased spacing but overall these configurations are highly detrimental for the formation of Sox/Oct complexes. The RF orientation is less obstructive. Whilst none of the sequences are bound with strong cooperativity, the RF+2 and RF+3 configurations include a number of sequences with weakly cooperative binding. Therefore, these configurations have the potential to support constructive dimer formation with functional relevance.

We next constructed models of the DNA-binding domains (DBDs) of Sox2 and Oct4 bound to the tested orientations with 0bp spacers. The propensity for steric clashes counteracting efficient assembly on these elements was estimated by modeling and energy minimization while recording a per-atom clash score (Figure 3C). As expected, no clashed are observed for the FF0 element. Here, helix 3 of the Sox2 HMG box is engaged in several favorable interactions with the helix 1 of the POU specific domain of Oct4 (25,28,46). When the half-site orientation is modified, this interface is lost and other portions of the domains are juxtaposed. When modeled on the FR0 element, the C-terminal tail of the HMG box (residues 77–81) penetrates the POUs subdomain and a number of additional clashes are seen for basic amino acids between the core of HMG box and the POU_{HD} domain. Likewise, in the RR configuration, residues of helix 3 of the HMG (in particular residues 67 and 68) clash severely with the main chain of the POU_{HD}. In the RF configuration the POU_S is brought near to helix1 of the HMG box, but no obvious protein-protein clashes are seen in the RF model. However, some POU residues at the protein–DNA interface exhibit high clash scores presumably as a consequence of local structural adjustments ac-

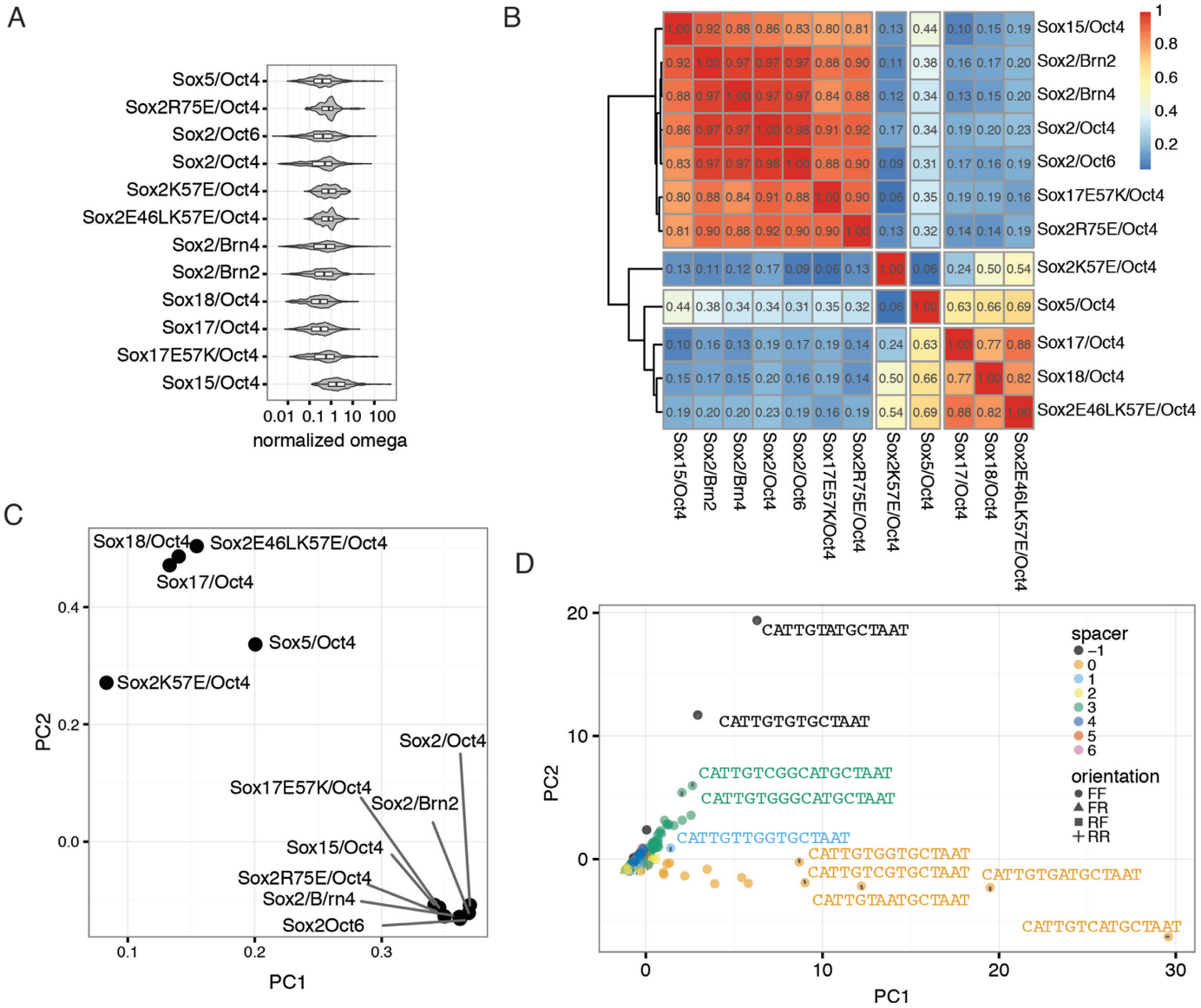

Figure 2. Coop-seq captures a broad range of TF cooperativties leading to the identification of specific heterodimer clusters and underlying sequence determinants. (**A**) Violin with overlaid box-and whisker plots showing that the cooperativity for the studied 12 Sox/Oct pairs varies over 3–4 orders of magnitude. (**B**) Correlation heatmap of the cooperativity factors for the 12 Sox/Oct pairs (rows and columns). Pearson correlation coefficients were hierarchically clustered. Fields are color coded by correlation coefficients. The main clusters are indicated with column/row spacing. (**C**) PCA variable loadings PC1 and PC2 are shown for the 12 Sox/Oct pairs. (**D**) PCA scores of individual data points projected along PC1 and PC2. Data points are color-coded by element spacing and the four orientations are mapped to different symbols. Selected sequences that explain most of the dataset's variance are shown. For each spacer/orientation combination several sequence were analysed due to the randomization of selected positions (Figure 1C) leading to multiple data points with identical color/shape coding (i.e. black, orange and green circles for FF−1, FF0 and FF+1 in (D)).

companying energy minimization. More global structural rearrangements, which cannot be captured in our energy minimization protocol, could likely resolve those clashes. In sum, in accordance with Coop-seq data, structural modeling designates the RF as most permissive for the formation of Sox/Oct heterodimers amongst the three non-canonical orientations, whilst the FR and RR are highly detrimental.

Coop-seq classifies Sox transcription factors by their dimerization preference with Oct4

To further dissect the cooperativity pattern of the Sox/Oct pairs we generated a matrix of mean ω values for all half-site configurations including all sequences variants and replicates (Figure 4A). Cooperative binding energies (ln(ω)) were hierarchically clustered to visualize the global dependence of Sox/Oct pairs on the motif configuration. This analysis demonstrates that the canonical FF0 and the compressed FF−1 elements show the most diversity in Sox/Oct pairing. In accordance with our previous studies, Sox17/Oct4 has a high ω value on the FF−1 configuration but cooperates more weakly than Sox2/Oct4 on the 'canonical' FF0 element (Figure 4B) (24,26). FF3 and RF2-4 configurations support moderately cooperative binding for a

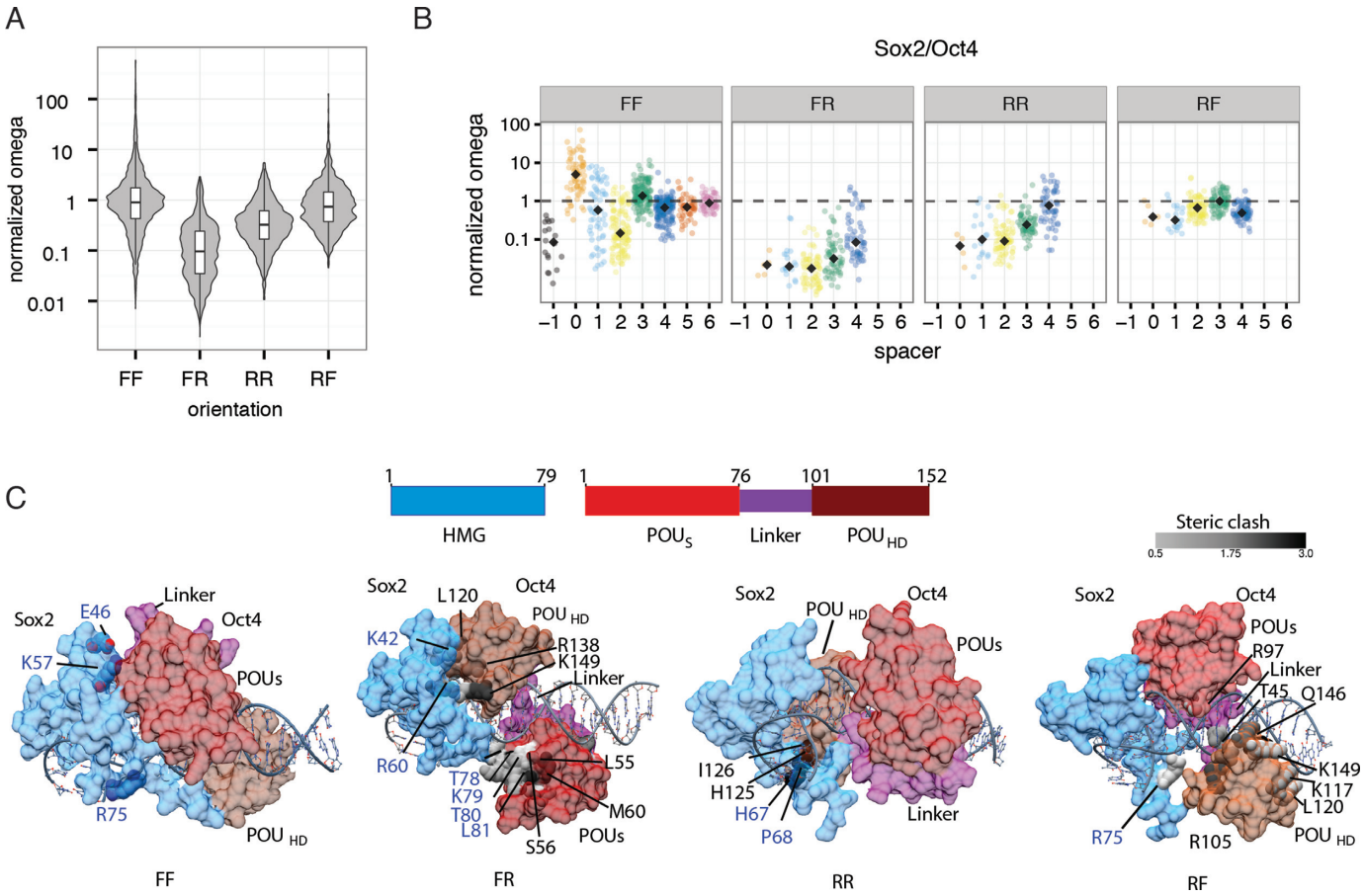

Figure 3. Half-site spacing and orientation dependence of the Sox/Oct cooperativity profiles. (**A**) Violin and box-and-whisker plots illustrating the distribution of cooperativities for all Coop-seq data for all the 12 Sox/Oct pairs as a function of half-site orientation. (**B**) The cooperativity for the Sox2/Oct4 heterodimer formation is shown for the four orientations as a function of half-site spacing. Colored dots represent sequence variants as wells as replicates and the gray diamond denotes the median. (**C**) Structural models of Sox2/Oct4 dimers on four composite DNA elements with zero base pair spacer. The Sox2 HMG box is shown light blue and the POU domain of Oct4 in red (POU specific subdomain, POU_S) or brown (POU homeodomain, POU_{HD}). The POU linker is colored magenta. Proteins are depicted as van-der-Waals and residues predicted to clash after energy minimization are shown with space-filling spheres and shaded according to the severity of the clash. Residues involved in clashes are marked with blue labels (Sox2) or black labels (Oct4). The DNA backbone is shown as ribbon. Sox residues E46, K57 and R75 mutated to change the cooperativity profile are shown for the FF configuration.

number of Sox/Oct combinations. Two Sox factors can be set apart from the rest by their cooperativity profiles. First, Sox15 retains moderately cooperative binding even on elements highly detrimental for the majority of Sox/Oct pairs. Second, Sox5 is found to be the only factor with a strong cooperativity on both, FF0 as well as FF−1 elements. This behavior is in marked difference to Sox2 preferring the FF0 site, and Sox17 preferring the FF−1 site (Figure 4B). To ask whether this behavior is mirrored in genome-wide binding studies we inspected ChIP-seq peaks of Sox2, Sox17 (14) and of Sox6 (47). Sox6 is highly similar to Sox5 and both belong to the SoxD subgroup. Sox6 was chosen, as no Sox5 dataset is available from public databases. We performed word searches in the binding peaks using degenerate IUPAC strings for FF0 and FF−1 motifs. As expected, we found an excess of FF0 elements in Sox2 peaks and an excess of FF−1 elements in Sox17 (Figure 4C). In the 19355 reported Sox6 peaks, however, we counted nearly equal amounts of FF0 and FF−1 sites (5545 versus 5568). This is consistent with the Coop-seq data for Sox5 showing equivalent cooperativity on both elements.

Neural POU TFs and Oct4 heterodimerize with Sox2 in a similar fashion

The four class III POU factors comprise a subgroup of the 14-member POU family. Many POUIII TFs are expressed in ectodermal lineages particularly in neural progenitors and the central nervous system (48) (Supplementary Figure S1) and were found to functionally partner with Sox2 (49). Class III POU factors were part of defined TF cocktails, some of which contain Sox2, to directly differentiate neural stem cells and mature neurons (50,51). We were thus particularly interested to understand differential co-binding of POUIII factors with Sox2 in comparison to the Sox2/Oct4 heterodimer. As Sox2 regulates both, pluripotency as well as neural differentiation (23), a partner switch from Oct4 to POUIII proteins such as Brn2 and Oct6 could provide the mechanistic basis for these dual and seemingly contradictory roles. Therefore, our Coop-seq study included three mouse POUIII factors Brn2, Oct6 and Brn4 (Figures 2B, C and 4A). To our surprise, on all tested half-site configurations we observed near-identical cooperativity profiles for

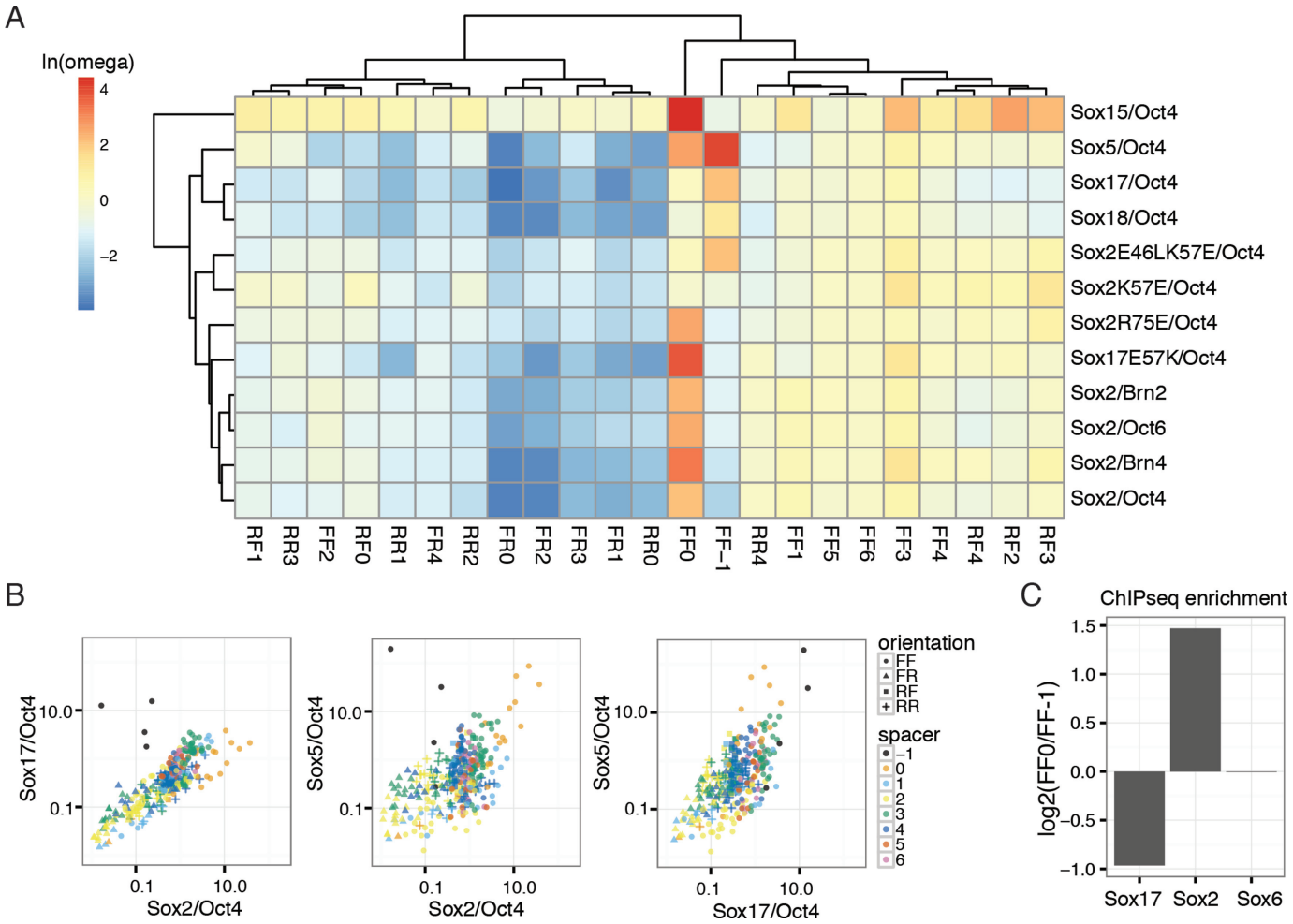

Figure 4. Sox factors associate with Oct4 in a DNA-element dependent fashion. (**A**) Hierarchically clustered heatmap of cooperativity factors for 23 element types (spacer/orientation combinations) and 12 Sox/Oct pairs. The natural log transformed mean cooperativity over all sequence variants and replicates per element type was used. The heatmap was prepared using the pheatmap R package with hierarchical clustering of rows and columns. (**B**) Pair-wise scatter plots of mean ω values for each analyzed sequence to highlight the differential binding of Sox17/Oct4, Sox2/Oct4 and Sox5/Oct4 to sequences with FF−1 and FF0 configuration (left panel). Data points are color-coded for half-site spacers and symbol-coded for orientations. As multiple sequence variants were used per spacer/orientation combination individual color/shape combinations occur multiple times (i.e. four black circles are present for the four FF−1 sequences studied (CATTGTNTGCTAAT). (**C**) Chip-seq peaks of Sox2, Sox17 (14) and Sox6 (47) were searched for the presence of 'compressed' FF−1 and 'canonical' FF0 sequences using the IUPAC strings FF0 = HWTTGWNATGYWWWD and FF−1 = HWTTGWATGYWWWD. Log2 transformed motif count ratios per dataset are plotted as barchart.

Oct4 and all three POUIII. This profile can be summarized by a very strong cooperativity on all FF0 sequences, moderate cooperativity on FF3 and a subset of FF1 sequences and competitive binding to FF−1 sequences (Figure 5A–C). Cluster analysis further illustrates the highly similar cooperativity profiles of all four POU proteins (Figures 2B and 4A). To rationalize this binding pattern, we constructed structural models of Sox2/Oct4 and Sox2/Oct6 dimers on the FF0 element (Figure 5D and E). Residues from the C-terminus of helix 1 of the POUs form the Sox2 interaction surface. In line with the near-identical cooperativity profiles the interaction interfaces are highly conserved. For example, residues 18, 12 and 23 are identical between the four proteins. However, Oct4 contains a threonine at position 22 that is replaced by a basic lysine in class III POU factors (Figure 5F). Importantly, only Oct6 constructs with the K22T mutations are able to maintain the pluripotency of mouse ESCs in a cellular complementation assay, an activity otherwise unique to the wild-type Oct4 protein (52). Yet, this amino acid does not modulate Sox/Oct dimerization to sequences interrogated in the present study leaving open the mechanistic basis for its functional relevance. It is possible that further randomizations within the target library can reveal differential heterodimerization profiles. Consistent with our study, Oct4 and Oct6 were reported to similarly bind to sox/oct elements with a canonical octamer sequence (ATGCTAAT) (52,53). However, only Oct4, but not Oct6, was found to heterodimerize with Sox2 on the *Utf1* element containing an atypical 'G' in the POU_{HD} half of the octamer (ATGCTAGT). A recent ChIP-seq study using Brn2 and Oct6 and neural progenitor cells provided further hints as to the mechanism how Oct4 and POUIII proteins can be set apart (54). *De novo* motif discovery in ChIP-seq enriched regions revealed the palindromic so-

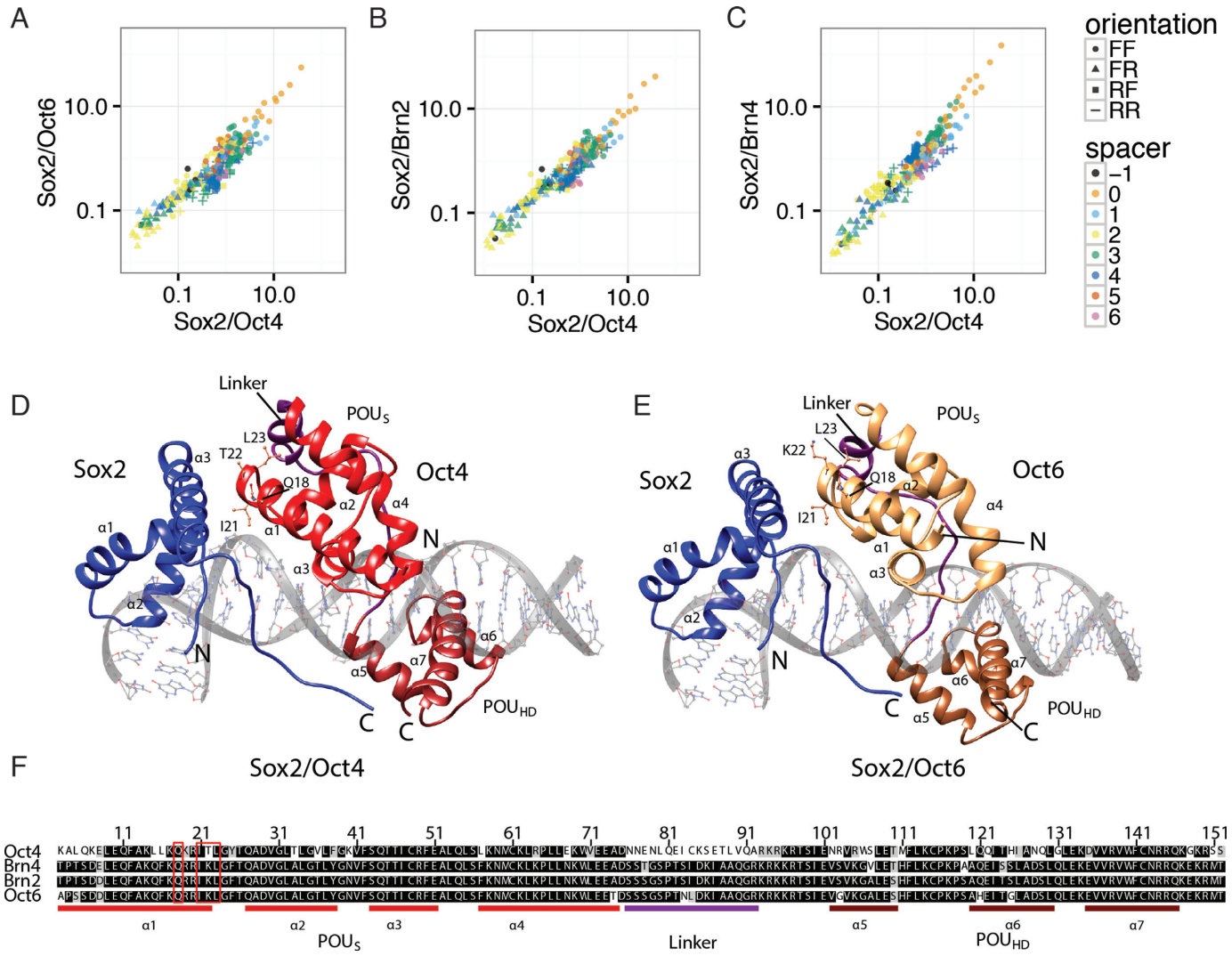

Figure 5. POU factors exhibit a near-identical cooperativity pattern with Sox2 despite directing contrasting cell fate decisions. (**A**–**C**) Pair-wise scatter plots of ω values to compare the cooperativity profile of the Sox2/Oct4 dimerization important during the maintenance and induction of pluripotency and the dimerization of Sox2 with the neural POU factors Oct6, Brn2 and Brn4. The mean of five replicate measurements for every sequence is plotted. Structural models of Sox2/Oct4 (**D**) and Sox2/Oct6 (**E**) bound to the 'canonical' FF0 element. Amino acids mediating protein-protein interactions are show as ball-and-sticks. (**F**) Alignment of the POU domains of mouse Oct4, Brn4, Brn2 and Oct6. Sox2 contact amino acids are colored red. Identical residues are on black background, conservative replacements on gray background and non-conservative replacements on white background.

called 'MORE' element (ATGCATATGCAT-like) as the top-scoring motifs for Oct6 and Brn2. Oct6 and Oct4 can form homodimers on the MORE sequence (55). By contrast, the canonical Sox/Oct element was reported as the top scoring motif in all published Oct4 ChIP-seq studies (11–14). Mistri at al. suggested that the DNA dependent Sox2/Oct4 heterodimer complex is formed more effectively than Sox2/POUIII complexes. In this model, the presence of Sox2 leads to a recruitment of Oct4 to Sox/Oct elements in pluripotent cells whilst in neural cells, where Sox2 is also strongly expressed but Oct4 is not, POUIII complexes remain bound to the MORE-like sequences as they are less likely to partner with Sox2. Yet, the reported apparent binding constant estimates in the absence or presence of Sox2 vary only subtly in this study (54). Therefore, as an alternative explanation, the differential propensity to homodimerise on MORE sequences could explain the motif preferences observed in ChIP-seq data. Indeed, we recently quantified the homodimerisation of Oct4 and other POU factors to the MORE element and found that Oct4 homodimerizes less effectively on this element with a cooperativity factor ∼15-fold lower than Brn2 and Oct6 (56). We identified and validated the amino acid responsible for these differences and showed that this site switches binding preferences for Sox/Oct and MORE elements *in vitro* and influences iPSC generation in a cellular context (56). Together, these data argue that reduced homodimerisation on the MORE contributes more strongly than enhanced heterodimerisation with Sox2 to the disparate genome engagement of Oct4 in comparison to POUIII TFs. Nevertheless, interrogating Sox/Oct heterodimerization by Coop-seq using libraries with additional site randomizations offers a tool to further clarify this point.

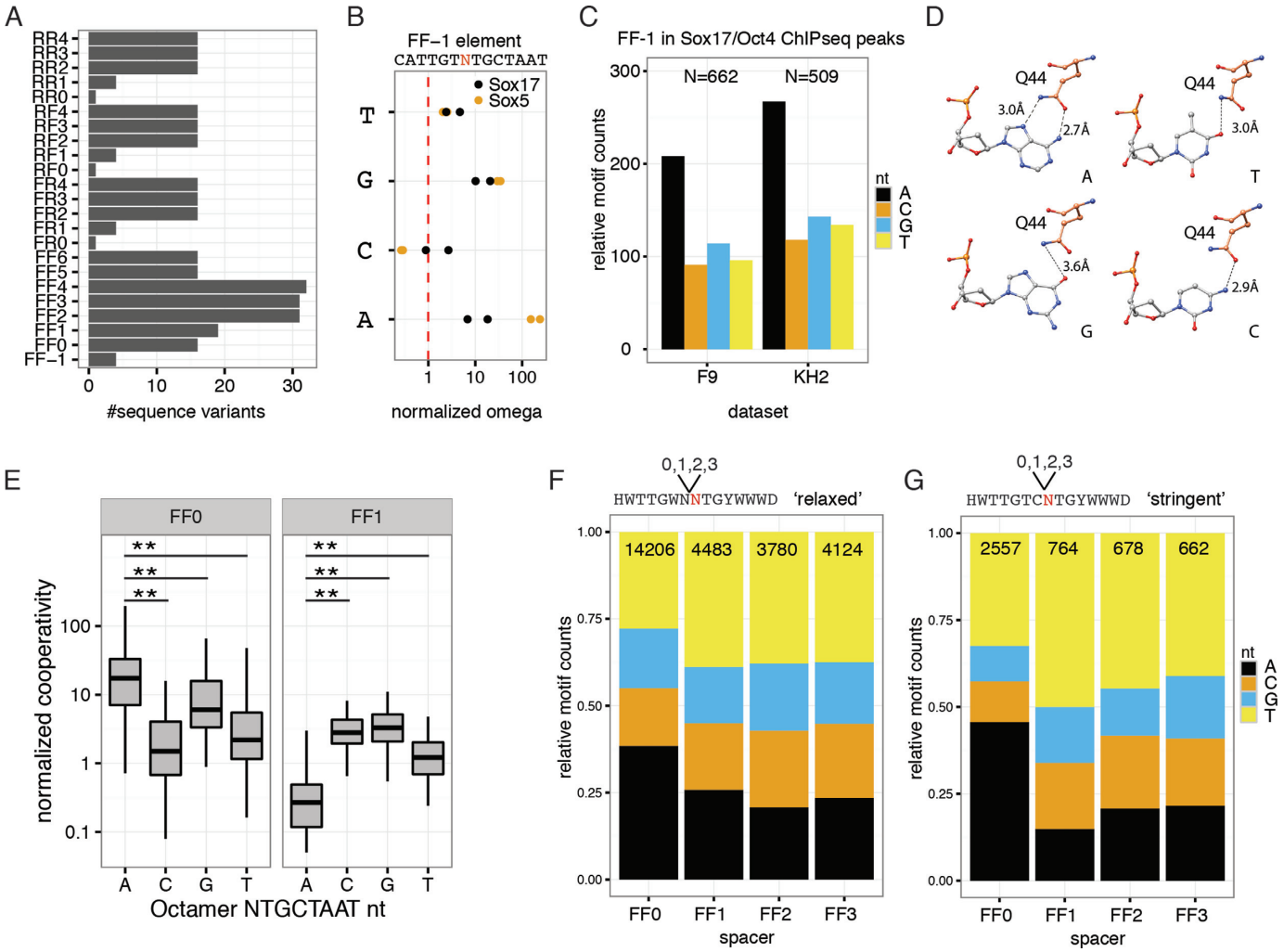

Figure 6. Nucleotide preferences in the octamer element change with altered half-sites spacing. (**A**) Number of sequence variants interrogated for the studied element types (orientation – spacer combinations). (**B**) Dependence of the normalized cooperativity (ω) on the identity of the randomized nucleotide is analyzed for the binding of Sox17 and Sox5 with Oct4 to the 'compressed' FF−1 sequence. (**C**) Genomic loci encoding FF−1 sequences co-bound by Sox17 and Oct4 in the KH2 mouse embryonic stem cell (mESC) line or retinoic acid (RA) treated embryonic carcinoma F9 cell line (14) were analyzed for nucleotide preferences at position 1 of the octamer sequence using the IUPAC string $H_1W_2T_3T_4G_5W_6[ACGT]_1T_2G_3Y_4W_5W_6W_7D_8$. (**D**) Contacts between position 1 of the octamer sequence and the Q44 of the POU-specific domain of Oct4. When Oct4 is bound to its cognate sequence (upper left) a favorable bidentate hydrogen bond is formed while this arrangement is disturbed upon replacement of the adenine. (**E**) Box-and-whisker plots to compare the dependence of the cooperativity on the identity of octamer nucleotide 1 for canonical FF0 and the FF1 elements. Coop-seq data for wild-type Sox2 binding with Oct4, Oct6, Brn2 and Brn4 were included for the analysis. Asteriks denote statistical significance from unpaired, two-side t-tests with *P*-values<0.01. (**F**, **G**) Genomic loci co-bound by Sox2 and Oct4 (57) were dissected for the relative preferences for position 1 octamer nucleotides as function of half-site spacing demonstrating that an 'A' is preferred for the canonical FF0 configurations but degenerate sites are preferred as the spacing increases. IUPAC strings used for the search are shown on top of the plots. Panel G uses a more stringent definition of the Sox site than panel F.

Coop-seq reveals a switch of base-preferences with altered half site spacing

The design of the Coop-seq libraries included randomized positions within spacer sequences but also at the edges of the core *Sox* and *Oct* motifs (Figures 1C and 6A). To explore whether Coop-seq can detect effects caused by such sequence variations, we focused on the sequences that randomize position 1 of the octamer binding sites (NTGCTAAT). For the FF−1 configuration, Sox17 and Sox5 prefer an A or a G at this position over a C or a T in agreement of with a strong enrichment of an A in Sox17/Oct4 co-bound sites in ChIP-seq studies (Figure 6B and C). This is in accordance with position-weight matrices obtained for Sox17/Oct4 co-bound sites (14). In structural models the A engages in favorable bi-dentate hydrogen bonds with the conserved Q44 of helix 3 of Oct4 (Figure 6D). As only the G exposes a hydrogen bond donor/acceptor combination within the major groove of the DNA, this constellation is altered when it is replaced likely necessitating structural adjustments to Q44. We next analyzed preference of Sox2/Oct dimers for nucleotides at position 1 of the octamer. As expected, an A is preferred on FF0 sequences followed by a G (Figure 6E). However, surprisingly, this preferences is reversed on the FF+1 element where the presence of an 'A' impedes positive cooperativity whereas C, G or T's permit dimer formation with a

median ω >1 (Figure 6E). To explore whether this switched nucleotide preference can also be detected in a cellular environment, we analyzed 13,129 genomic loci co-bound by Sox2/Oct4 in mouse ESCs (57). Indeed, for FF0 a higher proportion of sites possess an 'A' at position 1 of the octamer half-site but this proportion drops as the half site spacing increases (Figure 6F). This effect is more pronounced when the 3′-end of the Sox half site is further constrained (Figure 6G, TC instead of the less stringent WN where W is an A or T and N any nucleotide). Therefore, the FF+1 configuration might present a novel version of a Sox/Oct composite motif that has hitherto escaped detection, as it demands the replacement of the otherwise conserved 'A' at the beginning of the octamer. Possibly, replacing the A could cause Q44 to dislodge from its stable binding conformation (Figure 6D) and allow the POU to conformationally re-arrange to accommodate dimeric binding with Sox2.

CONCLUSION AND OUTLOOK

Coop-seq is a simple and powerful method providing accurate cooperativity parameters, which would otherwise be labor intensive and require 100–1000's of EMSA gels. The inherent noise of the assay, predominantly caused by PCR amplification and the deep sequencing reaction, can be effectively dealt with through replicate experiments, spacer randomization and appropriate normalization procedures. Aggregate results are consistent with dedicated EMSA assays on single sequences and composite motifs discovered from ChIP-seq experiments (14,26). Coop-seq could be scaled up to 1000 if not 10 000s sequences if the sequencing depth is increased. Still, the number of sequences sampled in parallel by Coop-seq is smaller than SELEX-seq and CAP-SELEX. However, Coop-seq provides biophysical parameters and a comprehensive assessment of the selected sequence set without complex modeling because all of the relevant states are assayed simultaneously. And because of the stringent EMSA conditions Coop-seq is unlikely to include false positive and false negative sequence elements. By contrast, SELEX-based approaches likely miss a large number of potentially functional configurations of binding elements and might lead to false positives because of enrichment artifacts. The data presented here provide a number of novel Sox/Oct dimer architectures warranting further functional interrogations. Those include the novel Sox15/Oct4 dimer on the RF+2 configurations, the Sox/POU dimers on the FF+1 element without 'A' at the beginning of the octamer or the strongly cooperative Sox5/Oct4 dimers on both 'canonical' and 'compressed' motifs.

Our set-up utilized protein domains purified to homogeneity. However, full length proteins and cellular extracts are expected to be suitable for Coop-seq as long as the individual microstates can be separated in native gels. Variants of Coop-seq could also be used to interrogate the homodimeric association of TFs which can require equally intricate configurations of binding elements (38,39). Moreover, more complex mixtures can be analyzed including more than two proteins or DNA elements with contrasting dimer configurations (such as heterodimers as well as homodimers in the same tube). Although for such complex mixtures with more than 4 protein/DNA microstates the mathematical formalism would need to be extended, as is used in MITOMI-based cooperativity experiments (37). The sequence libraries used in this study made only minimal modifications to the consensus binding sites. This likely explains why, for example, we did not detect differential binding for Oct4 and class III POU factors. TFs do not only partner via juxtaposed protein-protein interaction interfaces. Rather, they also mutually influence their binding energies indirectly mediated by the DNA as reported for Sox2/Pax6 and Sox2/Oct4 dimerization (58,59). Such a mechanism could explain the differential dimerization of Oct4 and Oct6 with Sox2 on a degenerate sequences as the one regulating the *Utf1* gene (52). Therefore, we expect Coop-seq to reveal further insights into the Sox/Oct partner code if the scope of the libraries is expanded or if libraries are designed that include native sequences identified in genome-wide binding studies.

DATA ACCESS

Raw reads for all of the experiments have been deposited in the NCBI short read archive under accession GSE85122.

Public datasets used for genomic data analysis:

Sox2/Oct4 (Figure 6F, G): GSE44288 (peaks called by us)
Sox6: GSM1717476 (peak files provided by the depositors used).
Sox2/Sox17/Oct4 (Figures 4C and 6C): GSE43275

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Basab Roy, Vikas Malik and Linlin Hou for helpful discussions and advice. Authors are indebted to Calista Keow Leng Ng (A*STAR, Singapore) for providing reagents.

FUNDING

National Institutes of Health (NIH) [HG000249 to G.D.S., GM101602 to J.J.H.]; 2013 MOST China-EU Science and Technology Cooperation Program [2013DFE33080 to R.J.]; National Natural Science Foundation of China [31471238]; 100 talent award of the Chinese Academy of Sciences and by Science and Technology Planning Projects of Guangdong Province, China [2014B030301058 and 2016A050503038]; Chinese Government Scholarship (CGS) and University of the Chinese Academy of Sciences (UCAS) (to Y.S.). Funding for open access charge: NIH [HG000249].

Conflict of interest statement. None declared.

REFERENCES

1. Jerabek,S., Merino,F., Scholer,H.R. and Cojocaru,V. (2014) OCT4: dynamic DNA binding pioneers stem cell pluripotency. *Biochim. Biophys. Acta*, **1839**, 138–154.
2. Schepers,G.E., Teasdale,R.D. and Koopman,P. (2002) Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families. *Dev. Cell*, **3**, 167–170.

3. Gubbay,J., Collignon,J., Koopman,P., Capel,B., Economou,A., Munsterberg,A., Vivian,N., Goodfellow,P. and Lovell-Badge,R. (1990) A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*, **346**, 245–250.
4. Harley,V.R., Lovell-Badge,R. and Goodfellow,P.N. (1994) Definition of a consensus DNA binding site for SRY. *Nucleic Acids Res.*, **22**, 1500–1501.
5. Werner,M.H., Huth,J.R., Gronenborn,A.M. and Clore,G.M. (1995) Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell*, **81**, 705–714.
6. Klaus,M., Prokoph,N., Girbig,M., Wang,X., Huang,Y.H., Srivastava,Y., Hou,L., Narasimhan,K., Kolatkar,P.R., Francois,M. *et al.* (2016) Structure and decoy-mediated inhibition of the SOX18/Prox1-DNA interaction. *Nucleic Acids Res.*, **44**, 3922–3935.
7. Herr,W., Sturm,R.A., Clerc,R.G., Corcoran,L.M., Baltimore,D., Sharp,P.A., Ingraham,H.A., Rosenfeld,M.G., Finney,M., Ruvkun,G. *et al.* (1988) The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. *Genes Dev.*, **2**, 1513–1516.
8. Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,C.O. (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell*, **77**, 21–32.
9. Hou,L., Srivastava,Y. and Jauch,R. (2016) Molecular basis for the genome engagement by Sox proteins. *Semin. Cell Dev. Biol.*, doi:10.1016/j.semcdb.2016.08.005.
10. 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.
11. Kim,J., Chu,J., Shen,X., Wang,J. and Orkin,S.H. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.
12. Kunarso,G., Chia,N.Y., Jeyakani,J., Hwang,C., Lu,X., Chan,Y.S., Ng,H.H. and Bourque,G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, **42**, 631–634.
13. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
14. 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.* (2013) Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. *EMBO J.*, **32**, 938–953.
15. Yuan,H., Corbi,N., Basilico,C. and Dailey,L. (1995) Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.*, **9**, 2635–2645.
16. 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.
17. Giorgetti,A., Montserrat,N., Rodriguez-Piza,I., Azqueta,C., Veiga,A. and Izpisua Belmonte,J.C. (2010) Generation of induced pluripotent stem cells from human cord blood cells with only two factors: Oct4 and Sox2. *Nat. Protoc.*, **5**, 811–820.
18. Huangfu,D., Osafune,K., Maehr,R., Guo,W., Eijkelenboom,A., Chen,S., Muhlestein,W. and Melton,D.A. (2008) Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat. Biotechnol.*, **26**, 1269–1275.
19. Yu,J., Vodyanik,M.A., Smuga-Otto,K., Antosiewicz-Bourget,J., Frane,J.L., Tian,S., Nie,J., Jonsdottir,G.A., Ruotti,V., Stewart,R. *et al.* (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science*, **318**, 1917–1920.
20. Kamachi,Y., Uchikawa,M. and Kondoh,H. (2000) Pairing SOX off: with partners in the regulation of embryonic development. *Trends Genet.*, **16**, 182–187.
21. Wilson,M. and Koopman,P. (2002) Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. *Curr. Opin. Genet. Dev.*, **12**, 441–446.
22. Irie,N., Weinberger,L., Tang,W.W., Kobayashi,T., Viukov,S., Manor,Y.S., Dietmann,S., Hanna,J.H. and Surani,M.A. (2014) SOX17 is a critical specifier of human primordial germ cell fate. *Cell*, **160**, 253–268.
23. Lodato,M.A., Ng,C.W., Wamstad,J.A., Cheng,A.W., Thai,K.K., Fraenkel,E., Jaenisch,R. and Boyer,L.A. (2013) SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. *PLoS Genet.*, **9**, e1003288.
24. 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.
25. Merino,F., Ng,C.K., 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.
26. 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.
27. Tapia,N., MacCarthy,C., Esch,D., Gabriele Marthaler,A., Tiemann,U., Arauzo-Bravo,M.J., Jauch,R., Cojocaru,V. and Scholer,H.R. (2015) Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency. *Sci. Rep.*, **5**, 13533.
28. Remenyi,A., Lins,K., Nissen,L.J., Reinbold,R., Scholer,H.R. and Wilmanns,M. (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.*, **17**, 2048–2059.
29. Kim,S., Broströmer,E., Xing,D., Jin,J., Chong,S., Ge,H., Wang,S., Gu,C., Yang,L., Gao,Y.Q. *et al.* (2013) Probing allostery through DNA. *Science*, **339**, 816–819.
30. Mirny,L.A. (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 22534–22539.
31. Nandi,S., Blais,A. and Ioshikhes,I. (2013) Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Res.*, **41**, 8822–8841.
32. Spivak,A.T. and Stormo,G.D. (2016) Combinatorial Cis-regulation in Saccharomyces species. *G3*, **6**, 653–667.
33. Jankowski,A., Szczurek,E., Jauch,R., Tiuryn,J. and Prabhakar,S. (2013) Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Res.*, **23**, 1307–1318.
34. Siggers,T., Chang,A.B., Teixeira,A., Wong,D., Williams,K.J., Ahmed,B., Ragoussis,J., Udalova,I.A., Smale,S.T. and Bulyk,M.L. (2012) Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. *Nat. Immunol.*, **13**, 95–102.
35. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
36. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
37. Isakova,A., Berset,Y., Hatzimanikatis,V. and Deplancke,B. (2016) Quantification of cooperativity in heterodimer-DNA binding improves the accuracy of binding specificity models. *J. Biol. Chem.*, **291**, 10293–10306.
38. BabuRajendran,N., Palasingam,P., Narasimhan,K., Sun,W., Prabhakar,S., Jauch,R. and Kolatkar,P.R. (2010) Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors. *Nucleic Acids Res.* **38**, 3477–3488.
39. Huang,Y.H., Jankowski,A., Cheah,K.S., Prabhakar,S. and Jauch,R. (2015) SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains. *Sci. Rep.*, **5**, 10398.
40. Zuo,Z. and Stormo,G.D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics*, **198**, 1329–1343.
41. Palmer,I. and Wingfield,P.T. (2004) Preparation and extraction of insoluble (inclusion-body) proteins from Escherichia coli. *Curr. Protoc. Protein Sci.*, doi:10.1002/0471140864.ps0603s38.

42. Nallamsetty,S. and Waugh,D.S. (2007) A generic protocol for the expression and purification of recombinant proteins in Escherichia coli using a combinatorial His6-maltose binding protein fusion tag. *Nature Protocols*, **2**, 383–391.
43. Jauch,R., Choo,S.H., Ng,C.K. and Kolatkar,P.R. (2011) Crystal structure of the dimeric Oct6 (POU3f1) POU domain bound to palindromic MORE DNA. *Proteins*, **79**, 674–677.
44. Arnold,K., Bordoli,L., Kopp,J. and Schwede,T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.
45. Stormo,G.D., Zuo,Z. and Chang,Y.K. (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief. Funct. Genomics*, **14**, 30–38.
46. Williams,D.C. Jr, Cai,M. and Clore,G.M. (2004) Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.*, **279**, 1449–1457.
47. Liu,C.F. and Lefebvre,V. (2015) The transcription factors SOX9 and SOX5/SOX6 cooperate genome-wide through super-enhancers to drive chondrogenesis. *Nucleic Acids Res.*, **43**, 8183–8203.
48. Suzuki,N., Rohdewohld,H., Neuman,T., Gruss,P. and Scholer,H.R. (1990) Oct-6: a POU transcription factor expressed in embryonal stem cells and in the developing brain. *EMBO J.*, **9**, 3723–3732.
49. Tanaka,S., Kamachi,Y., Tanouchi,A., Hamada,H., Jing,N. and Kondoh,H. (2004) Interplay of SOX and POU factors in regulation of the Nestin gene in neural primordial cells. *Mol. Cell. Biol.*, **24**, 8834–8846.
50. Han,D.W., Tapia,N., Hermann,A., Hemmer,K., Hoing,S., Arauzo-Bravo,M.J., Zaehres,H., Wu,G., Frank,S., Moritz,S. *et al.* (2012) Direct reprogramming of fibroblasts into neural stem cells by defined factors. *Cell Stem Cell*, **10**, 465–472.
51. Vierbuchen,T., Ostermeier,A., Pang,Z.P., Kokubu,Y., Sudhof,T.C. and Wernig,M. (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, **463**, 1035–1041.
52. Nishimoto,M., Miyagi,S., Yamagishi,T., Sakaguchi,T., Niwa,H., Muramatsu,M. and Okuda,A. (2005) Oct-3/4 maintains the proliferative embryonic stem cell state via specific binding to a variant octamer sequence in the regulatory region of the UTF1 locus. *Mol. Cell. Biol.*, **25**, 5084–5094.
53. Nishimoto,M., Fukushima,A., Okuda,A. and Muramatsu,M. (1999) The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. *Mol. Cell. Biol.*, **19**, 5453–5465.
54. Mistri,T.K., Devasia,A.G., Chu,L.T., Ng,W.P., Halbritter,F., Colby,D., Martynoga,B., Tomlinson,S.R., Chambers,I., Robson,P. *et al.* (2015) Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells. *EMBO Rep.*, **16**, 1177–1191.
55. Tomilin,A., Reményi,A., Lins,K., Bak,H., Leidel,S., Vriend,G., Wilmanns,M. and Schöler,H.R. (2000) Synergism with the coactivator OBF-1 (OCA-B, BOB-1) is mediated by a specific POU dimer configuration. *Cell*, **103**, 853–864.
56. Jerabek,S., Ng,C.K., Wu,G., Arauzo-Bravo,M.J., Kim,K.P., Esch,D., Malik,V., Chen,Y., Velychko,S., MacCarthy,C. *et al.* (2016) Changing POU dimerization preferences converts Oct6 into a pluripotency inducer. *EMBO Rep.*, doi:10.15252/embr.201642958.
57. 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.
58. Merino,F., Bouvier,B. and Cojocaru,V. (2015) Cooperative DNA recognition modulated by an interplay between protein-protein interactions and DNA-mediated allostery. *PLoS Comput. Biol.*, **11**, e1004287.
59. Narasimhan,K., Pillay,S., Huang,Y.H., Jayabal,S., Udayasuryan,B., Veerapandian,V., Kolatkar,P., Cojocaru,V., Pervushin,K. and Jauch,R. (2015) DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res.*, **43**, 1513–1528.