High-throughput biochemical analysis of in vivo location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes

Dean Tantin,[1,4] Matthew Gemberling,[2,4] Catherine Callister,[1] and William Fairbrother[2,3,5]

[1]*Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA;* [2]*MCB Department, Brown University, Providence, Rhode Island 02912, USA;* [3]*Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912, USA*

The transcription factor POU5F1 is a key regulator of embryonic stem (ES) cell pluripotency and a known oncoprotein. We have developed a novel high-throughput binding assay called MEGAshift (microarray evaluation of genomic aptamers by shift) that we use to pinpoint the exact location, affinity, and stoichiometry of the DNA–protein complexes identified by chromatin immunoprecipitation studies. We consider all genomic regions identified as POU5F1-ChIP–enriched in both human and mouse. Compared with regions that are ChIP-enriched in a single species, we find these regions more likely to be near actively transcribed genes in ES cells. We resynthesize these genomic regions as a pool of tiled 35-mers. This oligonucleotide pool is then assayed for binding to recombinant POU5F1 by gel shift. The degree of binding for each oligonucleotide is accurately measured on a custom oligonucleotide microarray. We explore the relationship between experimentally determined and computationally predicted binding strengths, find many novel functional combinations of POU5F1 half sites, and demonstrate efficient motif discovery by incorporating binding information into a motif finding algorithm. In addition to further refining location studies for transcription factors, this method holds promise for the high-throughput screening of promoters, SNP regions, and epigenetic modifications for factor binding.

[Supplemental material is available online at www.genome.org.]

Gene expression is mediated by the interaction of transcription factors with accessible locations in chromatin. Modulation of this accessibility and changes in the activity and composition of transcription factors are a major source of gene regulation during development. The transcription factor POU5F1 (also known as Oct4) has been implicated in maintaining embryonic stem (ES) cell pluripotency and also in reprogramming somatic cells to an ES cell fate (for review, see Pan et al. 2002; Wernig et al. 2007). POU5F1 was isolated from ES cells on the basis of its ability to bind an octamer sequence, ATGCAAAT (Scholer et al. 1989). It was later shown to be a principal factor in maintaining a stem cell state—a property that generated great interest in this transcription factor's target genes (Niwa et al. 2000). POU5F1 expression may also mark adult germline compartments and certain classes of tumors (Gidekel et al. 2003; Kehler et al. 2004; Atlasi et al. 2007). POU5F1's in vitro binding specificity has been determined by SELEX, a method of identifying high affinity binding sites from random sequence through iterative steps of binding selection and enrichment (Nishimoto et al. 2003). Weight matrices are calculated from an alignment of the selected sequences and used to score the "closeness" of real sequences to the high affinity sites. However these methods often leave questions about the in vivo relevance of the output sequences, as natural selection may not always favor the highest binding affinity sites. In addition, it has long been observed that other factors such as chromatin accessibility greatly limit the usefulness of in vitro binding specificities for predicting sites in vivo (for a more complete discussion of this phenomena, see Wasserman and Sandelin 2004).

More recently, new methods have been developed to measure interactions between transcription factors and chromatin. These methods, termed ChIP-chip and ChIP-PET, locate binding sites in vivo by immunoprecipitating the factor of interest after it has been cross-linked to chromosomal DNA (Orlando and Paro 1993). Binding regions are identified either by microarray (ChIP-chip) or by sequencing (ChIP-PET). Both of these techniques have been applied to the identification of POU5F1-bound regions in human and murine ES cells (Boyer et al. 2005; Loh et al. 2006).

While high-throughput localization studies such as ChIP-chip and ChIP-PET have revolutionized the field of transcription, they have several important limitations. Chemical cross-linking does not provide a quantitative measure and sometimes not even a measure of direct binding. The resolution is limited by the shearing size of the DNA prior to the immunoprecipitation, as well as microarray densities and sequencing depths (Buck and Lieb 2004). Most of the published location studies return regions that span ~1 kb of genomic space. The number and location of binding sites within these regions cannot immediately be determined. Recent advances in sequencing technology and the practice of size selecting the recovered DNA fragments have improved this situation, allowing for the inference of binding sites to a resolution of 50 base pairs (Johnson et al. 2007; Robertson et al. 2007). However, this approach remains costly and the resolution

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail Fairbrother@brown.edu; fax (401) 863-9653.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.072942.107.

is dependant on the number of binding sites in the genome, and so the stated resolution will vary. Furthermore, the inference that predicts a binding site from the distribution of sequenced tags becomes problematic when recognition elements cluster in closely spaced groups along the DNA. A more direct determination of POU5F1 binding sites within these regions will help identify variants that disrupt binding and also shed light on the mechanism of POU5F1 function. It has been shown that POU5F1 binding does not always enhance transcription. In some cases, POU5F1 binding is correlated with a repressed transcriptional state (Boyer et al. 2005). This duality is not uncommon for transcription factors and is probably explained by the local context of each site and the identity of neighboring factors on the DNA. High-resolution maps of transcription factor binding sites (TFBSs) in promoters are required to understand these nuances of transcription factor function. To date, functionally defined target gene regulation by POU5F1 binding has been described for only a handful of genes at base pair resolution. These include POU5F1 itself, FGF2, NANOG, SOX2, and SPP1 (osteopontin) (Ambrosetti et al. 1997; Botquin et al. 1998; Chew et al. 2005; Rodda et al. 2005).

In order to pinpoint the specific sites and quantify the strength of binding of a factor to its targets in vivo, the large volumes of output sequence returned by high-throughput ChIP methods will need to be interrogated by traditional means such as the electrophoretic mobility shift assays (EMSA). As these genomic location studies become increasingly utilized, the field will require high-throughput technology to identify binding sites within these genomic regions. One potential method is to use protein-binding microarrays (PBMs) to measure binding affinities of a labeled protein to double-stranded substrates arrayed on a glass slide. This technique has been used to identify binding specificities for transcription factors and can measure the degree to which a protein binds to a particular sequence (Mukherjee et al. 2004). While PBM is high-throughput, it lacks some of the flexibility and qualitative features of EMSA. For example, EMSA can distinguish single from multiple molecules of bound protein. EMSA is a highly quantitative, well-established method that can be performed in whole-cell extracts, allows for the physical isolation of the bound product, and does not require a microarray to analyze the result.

Here, we demonstrate the feasibility of a high-throughput EMSA with an analysis of POU5F1 binding capacity throughout the POU5F1-ChIP-enriched regions (Boyer et al. 2005). The MEGAshift method utilizes inexpensive commercial sources of solid-phase DNA synthesis to remake regions of interest as large pools of short oligonucleotides (see Discussion). The reverse strands of these oligonucleotides are also synthesized as probes onto a custom oligonucleotide microarray. The oligonucleotides are incubated with recombinant POU5F1 and the bound and unbound fraction analyzed by microarray. From this analysis, a binding affinity can be directly measured for each oligonucleotide and, so by proxy, for each window along the POU5F1-enriched chromosomal region.

Results

Pooled oligonucleotide design and experimental scheme

We started with genomic regions that were found to be occupied by POU5F1 in vivo using ChIP-chip or ChIP-PET in both human and mouse ES cells (Boyer et al. 2005; Loh et al. 2006). We reasoned that the intersection of these results would contain the highest-confidence set of POU5F1 binding sites that are conserved across these two species. While the original ChIP-PET study reported 88 POU5F1 targets shared by the human and the mouse results, only 19 genomic regions fit this criterion of overlapping binding regions (19 in either species = 38 gene regions total) (Supplemental Table S1). While POU5F1 is generally known as an activator in the stem cell state and is turned off during differentiation (Pan et al. 2002; data not shown), only 39% of POU5F1 targets are transcriptionally active in human ES cells. Among the subset of these targets that overlap with mouse ChIP regions, 73% were in close proximity to genes that were expressed in ES cells. In this regard, the subset chosen for study is significantly different (P-value < 0.002, $\chi^2 = 10$, d.f. = 1) than the overall data set from which they were derived either because (1) the overlapping conserved set has a lower false positive rate or (2) POU5F1 sites that function as enhancers are subject to stronger purifying selection than their silencer or otherwise nonenhancer counterparts.

The ChIP-enriched regions were analyzed in both human and mouse. The binding regions in human and mouse did not always align perfectly, and so to extend the region of comparison, the union of this overlap was synthesized in human (Fig. 1, step 1). Each of these 38 regions was then used to generate a contig of 35-mers tiled in 19-nucleotide increments across the genomic region enriched in the ChIP assay, creating a total of 2468 genomic aptamers(Fig. 1, step 2). Universal primers flanking the 35-mer were designed to enable PCR amplification of the library en masse. This library then represented an approximate twofold coverage of the POU5F1 regions in human and the orthologous regions in mouse and is capable of reporting binding sites with 19-nucleotide resolution. The library was then tested for POU5F1 binding by EMSA (Fig. 1, step 3) using the complex pool as a probe. The shifted fraction was isolated and analyzed either by sequencing or by hybridizing the sample to probes on a custom oligonucleotide microarray (Fig. 1, step 4). By differentially labeling the EMSA-selected and nonselected initial pool, enrichment can be derived from the red/green ratios of hybridization intensities on the microarray.

Sequential enrichment of POU5FI-binding activity from a complex pool

In order to isolate the fraction of the library that binds POU5F1, an EMSA was performed in the presence of purified recombinant human full-length POU5F1. To establish the appropriately discriminating concentration of POU5F1 for this experiment and to control for possible binding contributions from the flanking primers, a canonical octamer site, derived from a human immunoglobulin heavy chain promoter, was amplified with a mutant control. At the optimized POU5F1 concentration, the wild-type control was saturated and the mutant probe showed trace amounts of binding (Fig. 2A, lanes 1–4). The radiolabeled oligonucleotide library represents 2468 different genomic 35-mer windows flanked by the universal primer pair but migrated as a single band in the polyacrylamide gel (Fig. 2A, lane 5) with no appreciable shift when incubated with recombinant POU5F1 (Fig. 2A, lane 6). The region of the gel where the POU5F1 shift would be expected to migrate (as determined by the positive control) was excised, reamplified, and used to reprobe POU5F1 in round 2 of the selection (Fig. 2A, lanes 7, 8). In round 2, an appreciable signal, consistent with an POU5F1-bound probe, was detected. To determine if this fraction also bound POU5F1 in whole-cell extract, the EMSA was repeated in whole-cell lysate derived from

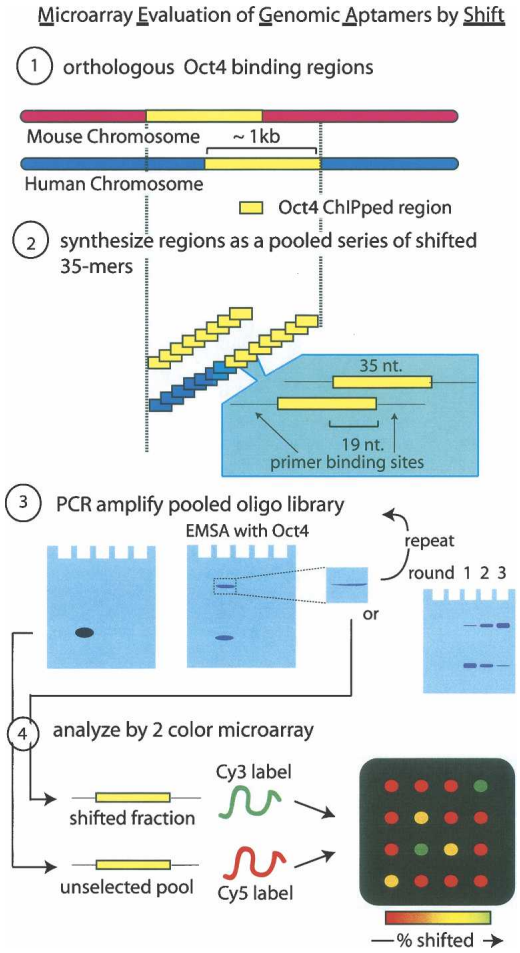

Figure 1. MEGAshift protocol. (Step *1*) All orthologous genomic regions enriched in both human and mouse POU5F1 (Oct4) ChIP experiments were aligned and resynthesized (step *2*) as a tiled contig of 35-mers flanked by universal primer binding sites. The human genomic region was extended to cover the union of this overlap. (Step *3*) This pool was amplified with labeled primers migrates as a single band and was then used in an EMSA activity with recombinant POU5F1. The shifted band was excised, reamplified, and either reshifted or analyzed by cloning or microarray. (Step *4*) Microarray analysis. Shifted and unselected fraction were reamplified and the T7 containing template used to generate Cy3 (shifted) or Cy5 (unselected) targets for the custom oligonucleotidearray.

ES cells using the enriched fraction from round 2 as a probe (Fig. 2B, lane 8).

Because of the tendency of POU5F1 to form various hetero and homo complexes, probes containing octamer sequences have been observed to display complex shifting patterns in extract (Remenyi et al. 2001; Remenyi et al. 2003). In this experiment, the wild-type probe displayed at least three shifted products that could represent POU5F1-containing complexes. While preincubating the extract with antibodies against the closely related POU2F1 had no effect on the formation of these complexes (Fig. 2B, lane 4), POU5F1-specific antibodies decreased the intensity of two of these shifted products and greatly increased the intensity of the third (arrow in Fig. 2B, lane 3). These three complexes can be efficiently competed with unlabelled wild-type probe and were also lost when EMSA was performed with differentiated ES cells, which lack POU5F1 (data not shown). Both with the wild type and selected pool, the upper band was responsive to the preincubation with POU5F1 antibodies (Fig. 2B, lane 3 vs. 9). A band of similar mobility was detected in the EMSA performed with the unselected pool but was not responsive to POU5F1 antibody (Fig. 2B, lanes 6, 7). This demonstrates that the enrichment protocol performed with the recombinant bacterially expressed POU5F1 selected for a population of probes that, as a group, had increased affinity for endogenous POU5F1 derived from ES cells.

The selection with recombinant POU5F1 was repeated an additional round resulting in more enrichment. In this final round 3, a faint band corresponding to a probe bound by multiple POU5F1 molecules was detected. The corresponding region of round 2 was excised from the gel from round 2. While no band was visible in this region of the gel, a product was amplified, and this pool resulted in enrichment in the additionally shifted molecules (see Fig. 2C, lanes 5, 6). This mixed population cannot be saturated within the concentration range of the protein required to saturate the positive control (Fig. 2D).

Recording the enrichment of 2468 genomic sites in the POU5F1 bound fraction of the pool

To determine the binding affinities of each of the 2468 oligonucleotides, we used a two-color labeling strategy in conjunction with an Agilent custom oligonucleotide microarray to compare the representation of each oligonucleotide in the shifted band to the enrichment of that oligonucleotide in the starting pool. The hybridization intensity of the selected targets was normalized to the hybridization of the pool by a process of multiplying each probe intensity in the selected channel by a constant such that the log of the ratio of selected/pool intensities (red/green spot ratios) summed to zero across all the probes on the microarray.

Initially all probes indicate a similar level of enrichment, but as the SELEX progresses, the tight distribution of enrichment scores gradually spread into progressively enriched and depleted subclasses of molecules (Fig. 3A). Sequencing of the initial pool verifies that the oligonucleotide synthesis protocol had proceeded with high fidelity and without any detectable enrichment bias (data not shown). We cloned 48 oligonucleotides from the POU5F1-enriched sets, and sequencing revealed that these consist of 39 unique clones where eight of these sequences were cloned multiple times (Supplemental Tables S2, S3). In addition to these eight clones, there were several other cases where a sequence was represented multiple times in overlapping clones (Fig. 4A). Presumably regions that were cloned multiple times from the selected pool were more highly enriched in the selected pool because they contain strong POU5F1 binding sites. To evaluate this prediction, the result of EMSAs performed with oligonucleotides cloned from the selected fraction, or the unselected fraction was compared to microarray enrichment (Fig. 3C,D). In this experiment, the pool was labeled with Cy5 and the selected fraction with Cy3 so oligonucleotides that are greenish in color are detected as enriched in the selected fraction (Fig. 3E). Oligonucleotides cloned from the selected pool both have a higher ranked enrichment (Fig. 3B, blue vs. red lines) and greater inclusion frequencies than oligonucleotides that were cloned from the unselected pool. An exception to this trend, unselected clone 4, was predicted by EMSA to be bound by POU5F1 but was cloned from the unselected fraction. This unexpected result is, however, consistent with the microarray estimate of enrichment; leading to the conclusion that lane 4 is a high affinity POU5F1

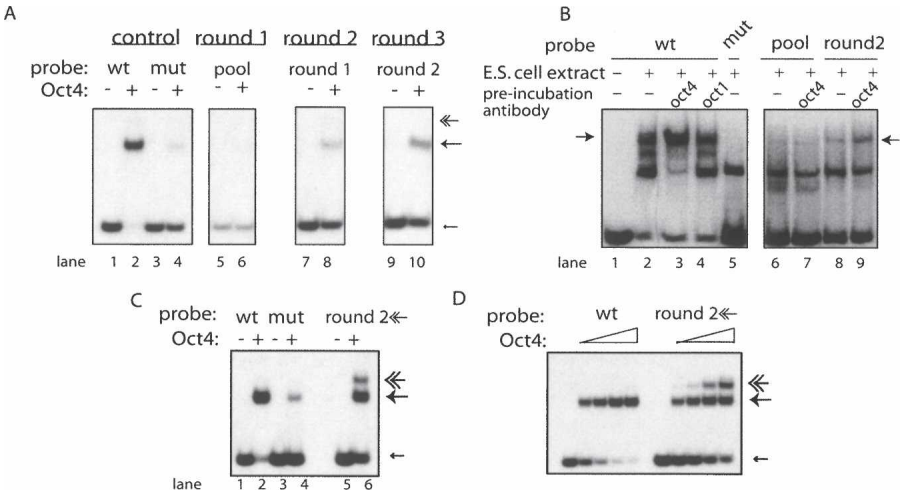

Figure 2. Enrichment for POU5F1 (Oct4) binding sites. (*A*) Perfect octamer, ATGCAAAT, containing oligonucleotide from the immunoglobulin heavy chain promoter ("wt" lanes *1,2*) and its octamer scrambled control ("mut" lanes *3,4*) were analyzed by EMSA with recombinant POU5F1(even number lanes) or no protein control (odd lanes). Mobility associated with singly bound oligonucleotides marked with an arrow; multiply bound oligonucleotides have been marked with a feathered arrow. Round 1 (lanes *5,6*) of the POU5F1 enrichment was performed with the synthetic oligonucleotide pool as a probe. The singly shifted fraction was excised, reamplified, and used as a probe in lanes *7,8*. The singly shifted fraction from round 2 was used as a probe in lanes *9,10*. (*B*) EMSA performed using J1ES cell extract. Extract was preincubated with POU5F1 (Oct4) antibody (lanes *3,7,9*) or with an antibody against a closely related POU2F1 (OCT1). Bands responsive to anti-POU5F1 antibodies are indicated with arrows. Lanes *6–9* display the oligonucleotide pool being shifted by ES extract. The starting oligonucleotide pool was used as probes for lanes *6,7*, while lanes *8,9* use selected sequences from Round 2 (*A*, lane *8*) of SELEX using the recombinant POU5F1. (*C*) Wild-type (wt), mutant, and the multiply shifted fraction of the oligonucleotide pool, excised (undetectable) from panel *A*, lane *8* (feathered arrow) were reamplified and used as a probe in an EMSA using recombinant POU5F1. (*D*) EMSA analysis performed using an increasing concentration gradient of recombinant POU5F1 protein on wild type and the multiply shifted fraction of the oligonucleotide pool.

binding sequence that happened to be cloned from the unselected pool (Fig. 3D, lane 4).

Distribution of POU5F1 binding sites in mammalian promoters

Custom genome tracks were written to facilitate the visual comparison of these cloning and enrichment results at each round of the SELEX experiment using the popular UCSC genome browser (Fig. 4). The region upstream of the *REST* gene represents an example of close agreement between the mouse ChIP-PET result and human ChIP-chip result. Oligonucleotides corresponding to the mouse and human were cloned multiple times from this region near the *REST* gene, also known as *NRSF*, and correspond well to the human site of maximal POU5F1 binding. This region is highly conserved (Fig. 4A) and also contains high-quality matches to POU2F1 weight matrices—motifs that are indistinguishable from the POU5F1 consensus sequences.

The region upstream of *GADD45G* represents an example of poor agreement between the mouse ChIP-PET results and human ChIP-chip result (Fig. 4B). Though the ChIPped regions contain little overlap, MEGAshift finds multiple sites that appear to be functionally conserved; i.e., enriched in both species throughout the SELEX experiment. The clustered distribution of binding sites that is common within the data appears to complicate the calling of enriched peaks in ChIP-chip analysis. For example, several closely spaced regions upstream of *GADD45G* are enriched in the ChIP-chip or ChIP-PET data. These regions overlap incompletely, yet according to MEGAshift, the most significant POU5F1 binding is occurring in a non-overlapping region. This oligonucleotide is located in a highly conserved block, comparable to an exonic coding sequence, and contains a predicted POU5F1 binding site. For these reasons, it is likely that this site is a bona fide POU5F1 binding site that was missed in the peak calling procedure.

Training binding models with enrichment data

To study the role of the POU5F1 consensus sequence in POU5F1 binding, all oligonucleotides were scored based on similarity

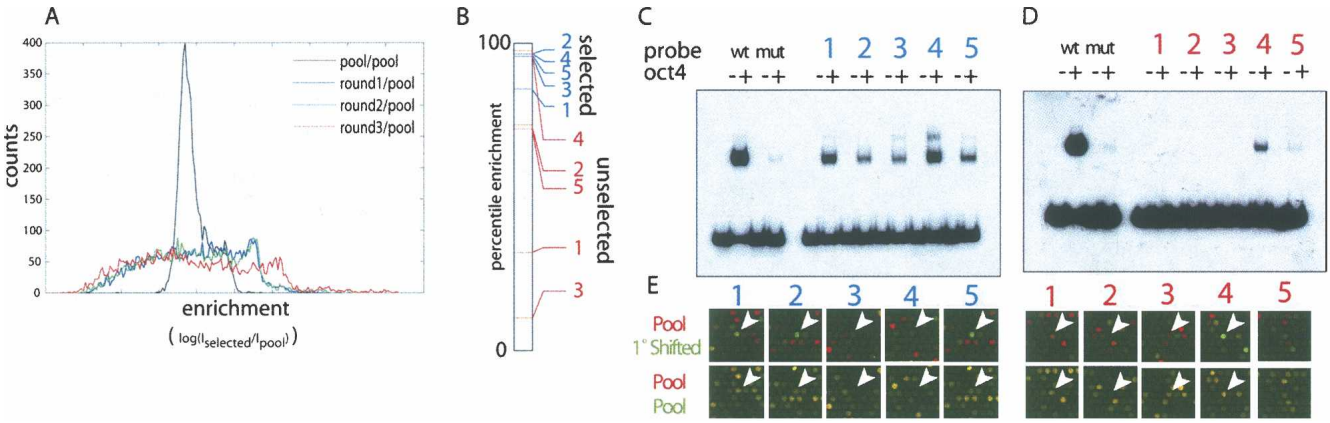

Figure 3. Changes in oligonucleotide enrichment throughout an POU5F1 SELEX experiment. (*A*) Enrichment scores for each round of SELEX were binned and graphed as a histogram. (*B*) Average enrichment scores were ranked with relevant oligonucleotides marked on the percentile bar. Gel shift assay was repeated for isolates cloned from selected (*C*) and unselected (*D*) fractions. (*E*) Microarray images corresponding to pool/pool and pool/round1 are drawn *below* the gel lane for each oligonucleotide.

with the SELEX-determined POU5F1 weight matrix (Supplemental Fig. 1S). The highest scoring window for each oligonucleotide was then compared with its enrichment value from the microarray. As expected, there is no initial bias for POU5F1 sites in the unselected pool's self comparison; however, throughout the course of SELEX, the higher scoring POU5F1 sites (lower values on the *Y*-axis; Fig. 5A) experience greater enrichment (a rightward migration on the *X*-axis; Fig. 5A) than the oligonucleotides without discernable POU5F1 sites (Fig. 5A, upper portion).

To determine whether incorporating the added information derived from the EMSA experiment increased the accuracy of motif prediction, we ranked the oligonucleotides according to enrichment in the singly shifted fraction of the pool (Round1) and multiply shifted fraction (multiple) (Supplemental Table S4). These ranked lists were then used to generate input sets for Gibbs sampler trials that searched for motifs of lengths 8–20 nucleotides long. Using the top 20 most-enriched probe signals (0.4% of the data), the Gibbs sampler converged on a motif that contained the consensus POU5F1 binding site (ATGCAAAT) in ~40% of the trials (Fig. 5B). This value peaked using ~3% of the data and decreased as a greater fraction of progressively less enriched data was added to the search space. Using this optimal amount of input data, we systematically explored the effect of motif length. While the POU5F1 consensus is 8 nucleotides long, the sampler converges to the consensus in only 50% of the runs when the motif length is set to eight, nine, or 10. In 97% of the cases where the sampler does not converge on the consensus, the output motif is a slightly truncated version of the consensus that extends to at least six positions of the octamer. We conclude that these sequences do not represent independent motifs. Indeed, both the singly shifted and the multiply shifted enrichment values include a wide range of motif lengths that return POU5F1 consensus sequences ~100% of the time (Fig. 5C; pictograms of all discovered motifs can be seen in Supplemental Fig. S1D). Judging by its performance with POU5F1, MEGAshift can be used to infer the binding specificity of a factor de novo. As MEGAshift can determine the identity of sequences present in both the singly or multiply bound state, it should be possible to learn sequence features that predispose a particular element to bind multiple molecules of POU5F1.

Sequence determinants of multimerization motifs

Plotting enrichment of the singly versus multiply bound fraction indicates that a sequence enriched in the singly bound fraction is

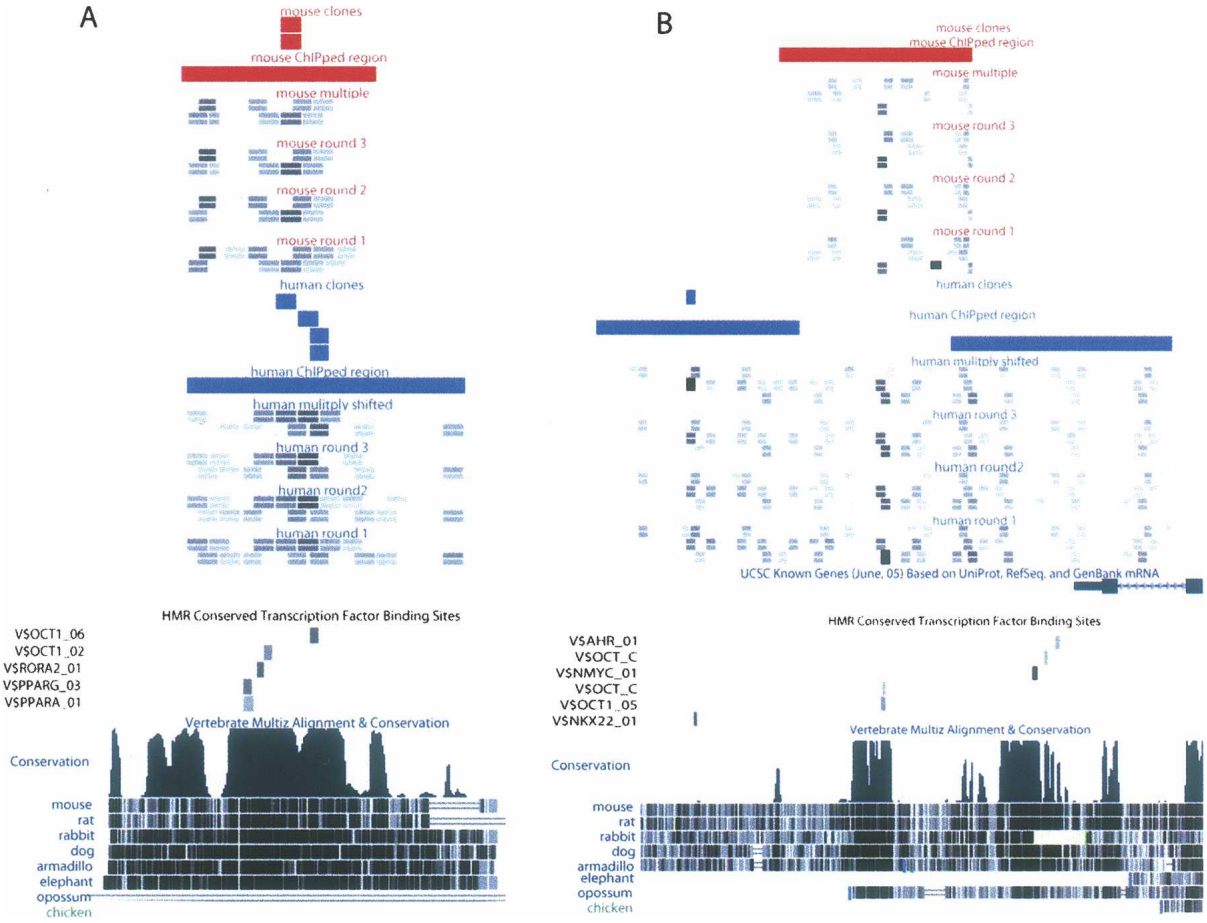

Figure 4. MEGAshift tracks for the UCSC Genome Browser. Two of the 19 genomic regions corresponding to REST (*A*) and *GADD45G* (*B*) have been displayed in custom UCSC genome browser tracks. Annotation is stacked vertically along the chromosomal coordinate axis (*X*-axis). Starting from the *top* and proceeding *down,* the mouse (red) sequences are annotated for the following molecules: oligonucleotides cloned out of POU5F1 selected fraction(short, stacked bars). ChIP-PET regions (wide bars), normalized enrichment scores in grayscale for each duplicate probe pair for each microarray experiment (multiply bound, round 3, round 2, round 1). Enriched oligonucleotides are shaded darkly. Human (blue) is identical save ChIPped material is analyzed by microarray (ChIP-chip). Predicted OCT binding sites annotated *below.* Conservation determined by eight vertebrate BLASTZ alignments.

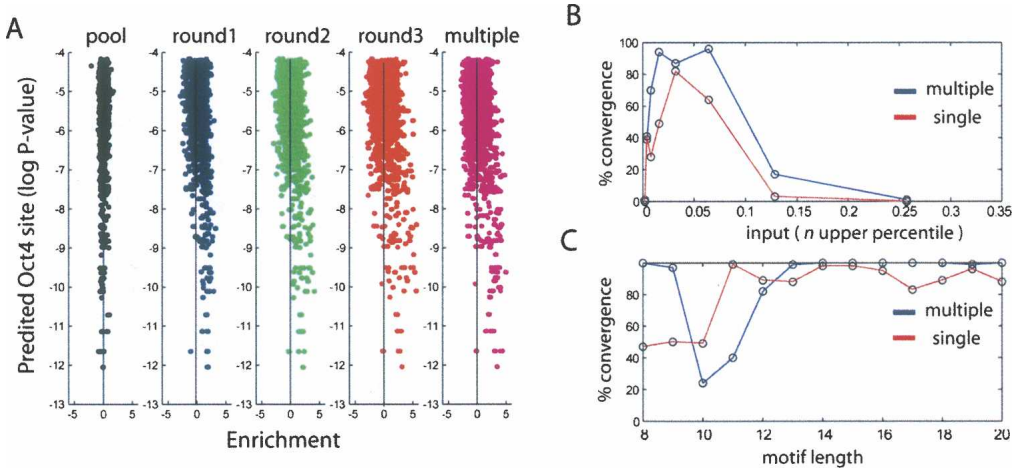

Figure 5. Comparison of octamer site prediction and POU5F1 binding. (*A*) POU5F1 sites were scored for each oligonucleotide as the log probability that a random sequence would fit the octamer binding model better than the highest scoring window (*Y*-axis) in the oligonucleotide and plotted against enrichment (*X*-axis). Vertical line represents mean enrichment for each experiment. (*B*) De novo motif identification was performed using Gibbs sampling trials with varying amounts of input that were ranked according to enrichment in round 1 (red) or the multiply bound fraction (blue). Successful trials that converged on motifs with the POU5F1 consensus (ATGCAAAT) were recorded on the *Y*-axis. (*C*) Using the top 3% of enriched oligonucleotides, the effect of motif length was examined.

more likely to also be enriched in the multiply bound fraction (Fig. 6A). Are there distinct sequence features that favor multimer formation? The Gibbs sampler converges on the core POU5F1 binding sequence in the multiply shifted pool, but the behavior of the length parameter suggests slightly different motif characteristics. As has been noted for other POU domain proteins, palindromic combinations of half sites have been known to support homo and heterodimer formation (Remenyi et al. 2001). Although these designed examples are informative, MEGAshift allows for the discovery of genomic sequences that predispose POU5F1 to bind as a multimer.

Changing the sampling parameters to include more than one binding model and restricting the oligonucleotide input for Gibbs sampling in such a way that retains the requirement for binding but emphasizes oligonucleotides that are bound as multimers returns three motifs that appear as chimeric combinations of POU5F1 half sites (Supplemental Fig. S1E). The full range of half site combinations found to be enriched in the multiply bound fraction (all points below the red line in Fig. 6A) relative to the singly bound fraction (all points above the green line) is noticeably more diverse than the combinations identified to date (Fig. 6B). In general, oligonucleotides that contained three or more half sites were enriched in the multiply shifted fraction Oligonucleotides that contained the ATGC half site also tended to be enriched in the multiply bound fraction. The well-studied palindromic half-site combinations represent only a minority of the total possible combinations observed in vivo in the selection of data present here (Fig. 6B) (Remenyi et al. 2001).

Discussion

These results demonstrate that the majority of nucleotide sequences identified by ChIP are not able to interact with recombinant POU5F1. Of the fraction of nucleotides that were able to bind, we demonstrate that these sequences can also interact with the POU5F1 in ES cell extracts. We identified numerous sequences with multiple paired and overlapping nonconsensus POU5F1-bound sequences, underscoring the difficulty of identifying biologically relevant sites using ChIP, SELEX, or in silico approaches by themselves.

These experiments serve as a bridge between low-resolution in vivo ChIP-chip results and a high-resolution molecular characterization of protein–nucleic acid interactions. Because MEGAshift is coupled to a readout of in vivo binding activity (ChIP), this technique represents a means of obtaining the best possible estimate of POU5F1 binding in ES cells. One intriguing feature of this work is the role of noncanonical binding sites that are comprised of various combinations of half sites. While this class of elements was not detected in SELEX experiments or in the original ChIP data, synthetic versions of these elements have been shown to facilitate POU5F1 binding as a dimer (Remenyi et al. 2001, 2003). These types of hybrid binding sites have been observed with other transcription factors, and MEGAshift offers a powerful discovery tool to characterize alternate modes of binding for these novel classes of sites (Wei et al. 2006; Johnson et al. 2007).

The finding that native POU5F1 binding sites frequently adopt complex multimeric configurations raises the interesting possibility of gene regulation either through induction of specific dimer assemblages or through differential transcriptional activity of particular dimer configurations. Evidence for the latter possibility comes from the finding that POU domain protein dimers formed on a consensus MORE site (ATGCATATGCAT) cannot recruit the transcriptional coactivator POU2AF1/OCA-B, whereas dimers formed using a PORE sequence (ATTTGAAATGCAAAT) are able to recruit POU2AF1 (Tomilin et al. 2000). Evidence for the former possibility comes from our recent findings that sequence-specific dimerization can be induced following stress through specific post-translational modifications of conserved residues in the POU DNA binding domain (data not shown).

Another application is to discover or refine binding motifs based on direct evidence and real sequence. Choosing the appropriate parameters and using a MEGAshift-ranked input leads to a complete convergence of motif finders on the octamer sequence. POU5F1 has a known binding motif, but this result demonstrates

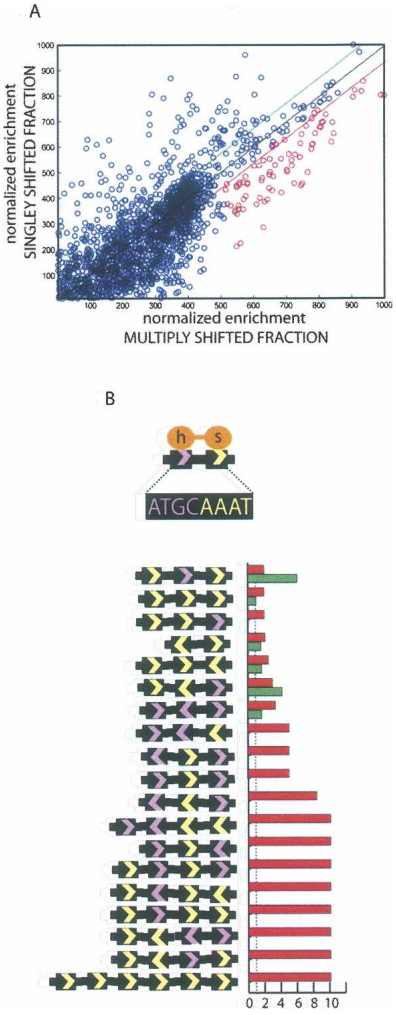

Figure 6. De novo motif identification. (*A*) For each oligo, singly bound (*Y*-axis) enrichment values were plotted against multiply bound enrichment (*X*-axis) values. POU5F1 contains two POU domains that recognize a bipartite signal as diagrammed in *B*. Half sites (ATGC, GCAT, AAAT, and ATTT) are counted in the entire set of oligonucleotidess, the set biased toward the singly bound state (above green line in *A*), and the set biased toward multiply bound (below the red line in *A*). Each permutation of half sites with more than twofold relative risk of being found in the multibound state versus the entire set is graphed. Red histogram bars mark relative risk (RR) of particular combination occurring in the multiply shifted fraction. Green bars for singly shifted fraction. Both measures are relative to the entire set and the blue dashed line marks zero enrichment (RR = 1).

that binding motifs could be identified from unknown complexes that form on oligonucleotide pools in whole-cell extracts. If upstream promoters of coordinately regulated genes were used for the design of oligonucleotide pools, these complexes could be mapped relative to each other, allowing for the definition of regulatory modules. Finally, sequences identified on the basis of one parameter (such as POU5F1 binding) can be tested for other attributes, such as general sequence composition and neighboring TFBSs.

One significant benefit of MEGAshift is the ability to specify the sequences of the oligonucleotides that go on to be assayed for binding. This assay could be used to discover candidate functional SNPs. Mutations or polymorphisms could be engineered into specific positions in an oligonucleotide such that both allelic versions of the oligonucleotide are present in the pool. MEGAshift could then be used to assay the effect of sequence variation on a transcription factor, like POU5F1.

MEGAshift is not limited to binding assays. The general idea of a synthetic oligo library, a molecular selection, and a microarray based readout could be reconfigured into a variety of functional assays. For example, RNA stability could be assayed in extract. Any application that involves a pool of nucleic acid and a method of molecular selection could be amenable to this protocol. MEGAshift is an affordable means to biochemically characterize a large sequence pool. Oligonucleotide pools can be inexpensively synthesized from commercial sources (e.g., Atactic makes 4000 sequences for approximately $700) or even mechanically scoured from custom oligonucleotide microarray (e.g., Agilent microarray yields 240,000 sequences at a cost of $450). These pools represent a one-time expense as the oligo library can be propagated by PCR amplification in multiple experiments. The microarrays to read the output are also economical. Microarrays can be stripped and reused several times, and the results could easily be analyzed by sequencing if microarray scanners are unavailable (Guenther et al. 2007).

In conclusion, MEGAshift represents a hybrid of biochemical, genomic, and computational approaches applied to the question of binding specificity in gene expression. It is extremely versatile in creating reagents that can be shared, reused, and cloned from. In summary, this method brings high-throughput experimentation to laboratories that may not have the requisite equipment and resources to follow typical high-throughput protocols.

Methods

Library design, oligonucleotide synthesis, cloning, and sequencing

The tool liftOver (Kent et al. 2002) was used to map the mouse coordinates onto the human genome, and a Perl script was used to identify overlapping regions. The human sequence was extended to completely cover the mouse sequences. A complex pool of 2468 oligonucleotides was synthesized in picoarray microfluidic μParaflo devices. The pool was amplified en masse by 10 cycles of PCR and end-labeled using [γ-^{32}P]ATP. Each oligonucleotide was designed as a tiled genomic 35-mer flanked by the common sequences CCAGTAGATCTGCCA and ATGGAGTC CAGGTTG that were used as the universal primer binding pair.

Recombinant human POU5F1 and preparation of whole-cell extracts

GST-human POU5F1 bacterial expression vectors and protocols were supplied by Drs. Yehudit Bergman (Hebrew University, Israel) and Jungho Kim (Sogang University, South Korea). Briefly, an overnight culture of BL21-DE3 (Codon-plus, Stratagene) *Escherichia coli* was diluted 1:20 in LB, grown to OD_{660} 0.5, and induced for 4 h with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at 30°C. Cells were lysed using SoluLyse (Genlantis) in 50 mM Tris-Cl (pH 8.0), 1% NP-40, 2 mM ethylene-diaminetetraacetic acid (EDTA), 150 mM NaCl, 0.5 mM dithiothreitol (DTT), plus a protease inhibitor cocktail (Roche). The lysate was clarified by centrifugation and incubated with glutathione-Sepharose (GE Healthcare) for 30 min at 4°C. Sepharose beads were collected by centrifugation and washed three times with the lysis buffer. Washed beads were eluted with 20 mM glutathione

in lysis buffer. The purified GST-POU5F1 in the eluate was dialyzed into buffer D (20 mM HEPES at pH 7.9, 100 mM KCl, 0.1 mM EDTA, 20% glycerol, 1 mM DTT, 0.5mM phenylmethylsulphonyl fluoride [PMSF]).

Whole-cell extracts were obtained from J1 ES (male) undifferentiated and retinoic acid (RA) differentiated cells. Cells were pelleted, resuspended, and incubated in extraction buffer (200 mM KCL, 100 mM Tris at pH 8.0, 0.2 mM EDTA, 0.1% Igepal, 10% glycerol, 1 mM PMSF) for 50 min on ice. Cell debris was pelleted, and extracts were frozen using liquid N_2 and stored in the −80°C.

Electrophoretic mobility shift assays

Oligonucleotides were prepared for EMSA by end-labeling PCR products with [γ-^{32}P]ATP. Samples were prepared in 20 µL (0.6× buffer D, 50 ng/µL Poly dI·dC, 1 µg/µL BSA, 1 mM DTT, 20 ng of probe). Samples were incubated for 30 min at room temperature. Native 4% polyacrylamide gels (29:1 acrylamide:bisacrylamide, 1% glycerol, 0.5× TBE) were prerun for 1 h at 80 V; samples were loaded and run for 1.75 h at 80V.

Hybridization and microarrays

Custom oligonucleotide microarrays (8 × 15 K) were produced by Agilent Technologies Inc. with default parameters. Probes were designed (in duplicate) to be the exact reverse compliment of the oligonucleotides in the library. Microarrays were hybridized following a modified version of the Agilent two-color microarray-based gene expression analysis protocol. Microarrays were hybridized for 3 h at 50°C and then washed for 1 min with 2× SSC with 0.2% SDS, for two 1-min washes with 1× SSC, and finally for one 10-sec wash with 95% EtOH. Microarrays were then centrifuged dry and scanned using a GenePix 4000B scanner from Molecular Devices. RNA probes were produced and labeled with Cy3 and Cy5 using the MEGAshortscript High-Yield Transcription kit (Ambion) after appending a T7 promoter to the oligonucleotides.

Cell culture

Differentiation of J1 ES cells occurred in the presence of 10^{-7} M (or 100 nM) retinoic acid (RA) over a 16-d period. ES cells were cultured in DMEM + HEPES supplemented with 1 mM glutamine, 1 mM sodium pyruvate, and 1 mM MEM nonessential amino acids (Invitrogen) plus 15% ES cell-qualified heat-inactivated fetal bovine serum (HyClone), 50 µM 2-mercaptoethanol (Sigma), and leukemia inhibitory factor (LIF/ESGRO, Chemicon).

Web resources

Genome browser snapshots of all the gene loci with POU5F1 binding data can be viewed at http://fairbrother.biomed.brown.edu/data/POU5F1. The raw data from the microarray experiments are stored as text files on the server, as is legend diagramming each experiment. Custom tracks will be submitted to the UCSC Genome browser and are also available for download.

Acknowledgments

We thank R. Freiman for invaluable advice with experiments and a critical reading of the manuscript. We also thank the anonymous referees for useful comments that improved the quality of the paper; Y. Bergman and J. Kim for providing recombinant POU5F1 vectors and purification protocols; C. Murtaugh and M. Capecchi for ES cell lines, and D. Stillman, T. Formosa, and C. Murtaugh for allowing us to use reagents and equipment. We thank members of the Tantin and Fairbrother laboratories for useful discussions and technical suggestions. This work was partially supported by a Richard Salomon Award (W.F.) and a March of Dimes/Basil O'Connor Award (D.T.).

References

Ambrosetti, D.C., Basilico, C., and Dailey, L. 1997. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.* **17:** 6321–6329.

Atlasi, Y., Mowla, S.J., Ziaee, S.A., and Bahrami, A.R. 2007. OCT-4, an embryonic stem cell marker, is highly expressed in bladder cancer. *Int. J. Cancer* **120:** 1598–1602.

Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M.K., Vriend, G., and Scholer, H.R. 1998. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. *Genes & Dev.* **12:** 2073–2090.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

Buck, M.J. and Lieb, J.D. 2004. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83:** 349–360.

Chew, J.L., Loh, Y.H., Zhang, W., Chen, X., Tam, W.L., Yeap, L.S., Li, P., Ang, Y.S., Lim, B., Robson, P., et al. 2005. Reciprocal transcriptional regulation of POU5F1 and Sox2 via the POU5F1/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.* **25:** 6031–6046.

Gidekel, S., Pizov, G., Bergman, Y., and Pikarsky, E. 2003. Oct-3/4 is a dose-dependent oncogenic fate determinant. *Cancer Cell* **4:** 361–370.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kehler, J., Tolkunova, E., Koschorz, B., Pesce, M., Gentile, L., Boiani, M., Lomeli, H., Nagy, A., McLaughlin, K.J., Scholer, H.R., et al. 2004. Oct4 is required for primordial germ cell survival. *EMBO Rep.* **5:** 1078–1083.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

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.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36:** 1331–1339.

Nishimoto, M., Miyagi, S., Katayanagi, T., Tomioka, M., Muramatsu, M., and Okuda, A. 2003. The embryonic Octamer factor 3/4 displays distinct DNA binding specificity from those of other Octamer factors. *Biochem. Biophys. Res. Commun.* **302:** 581–586.

Niwa, H., Miyazaki, J., and Smith, A.G. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.* **24:** 372–376.

Orlando, V. and Paro, R. 1993. Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* **75:** 1187–1198.

Pan, G.J., Chang, Z.Y., Scholer, H.R., and Pei, D. 2002. Stem cell pluripotency and transcription factor Oct4. *Cell Res.* **12:** 321–329.

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.

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.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4:** 651–657.

Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. 2005. Transcriptional regulation of nanog by OCT4 and

SOX2. *J. Biol. Chem.* **280:** 24731–24737.
Scholer, H.R., Balling, R., Hatzopoulos, A.K., Suzuki, N., and Gruss, P. 1989. Octamer binding proteins confer transcriptional activity in early mouse embryogenesis. *EMBO J.* **8:** 2551–2557.
Tomilin, A., Remenyi, A., Lins, K., Bak, H., Leidel, S., Vriend, G., Wilmanns, M., and Scholer, 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.
Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5:** 276–287.
Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.
Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448:** 318–324.

Received October 17, 2007; accepted in revised form January 16, 2008.