Combinatorial binding of transcription factors in the pluripotency control regions of the genome.

Luciana Ferraris, Allan Stewart, Jinsuk Kang, Alec DeSimone, Matthew Gemberling, Dean Tantin and Will Fairbrother
Center for Computational Molecular Biology, Center for Proteomics and Genomics. Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence 02912

Supplementary Information

Figures

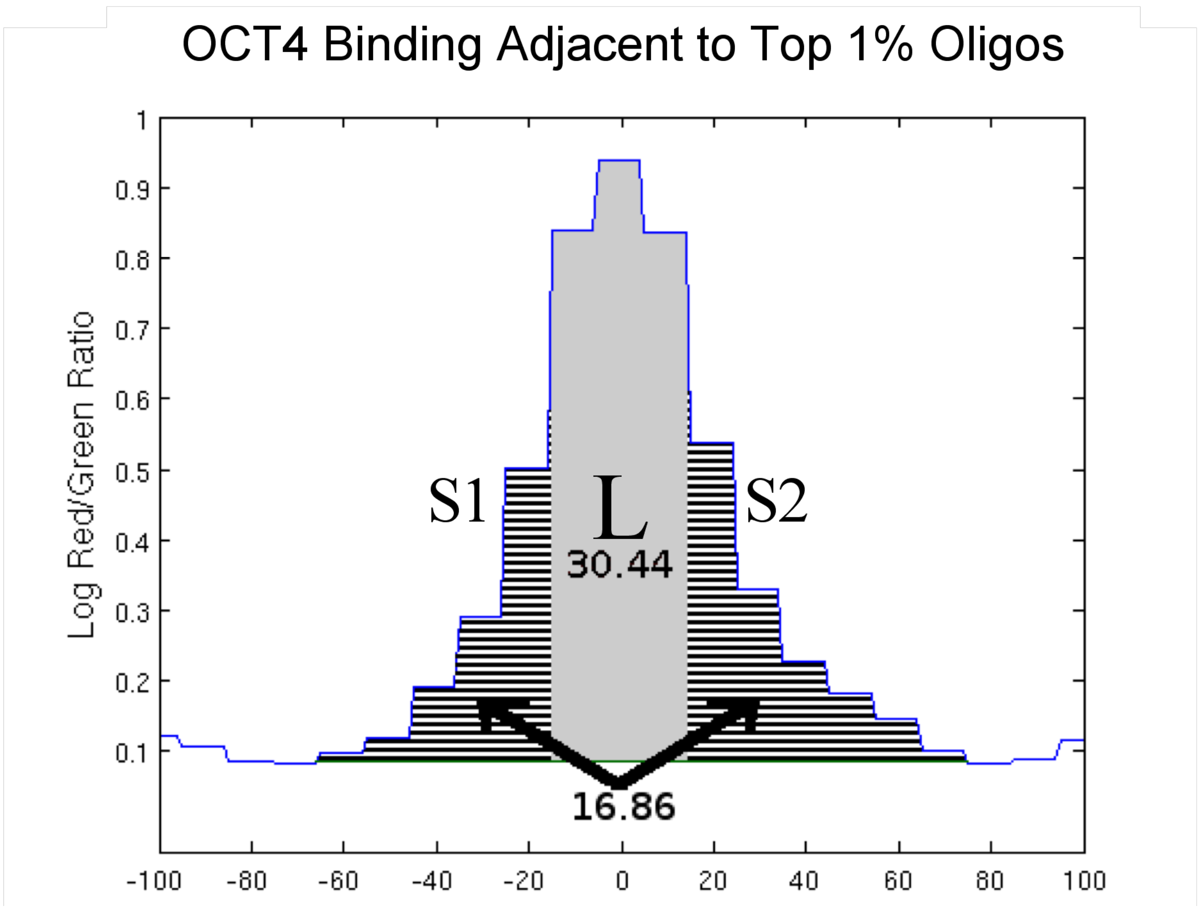


Figure S1, related to Figure 3: Example of Wide Binding in POU5F1 (formerly known as OCT4). We computed an average log ratio of binding for every position in the genome within 100bp upstream or 100bp downstream of the center of POU5F1 enriched oligos (top 1% enrichment). We plotted this curve and computed the area under the curve (and above a background rate of enrichment) for the oligo ligand and the adjacent regions. See Table S2 for data in all of the transcription factor enrichments.

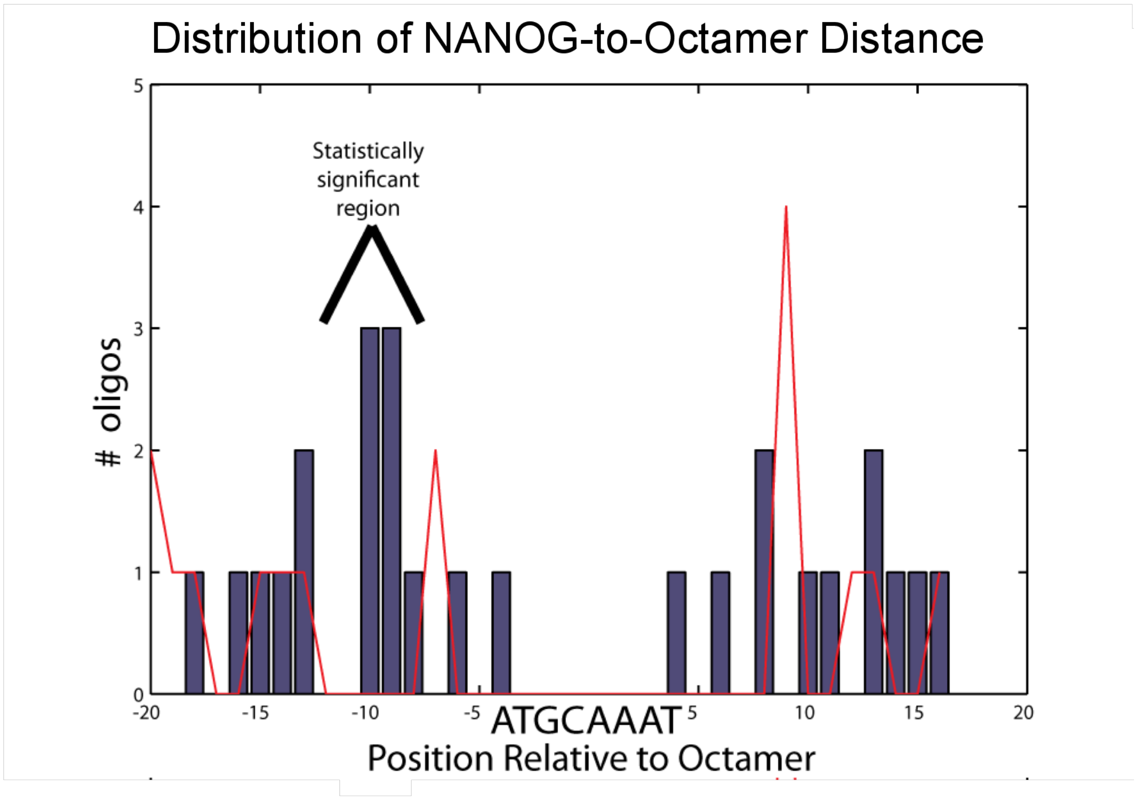


Figure S2, related to Figure 6: NANOG-to-Octamer Distance in top 5% NANOG enrichment relative to bottom 25% enrichment. The blue bars represent the position of the center of NANOG recognition elements (CATT) in the top 5% enrichment for NANOG binding. The red line plots the analogous distribution in the bottom 25% enrichment. We find that there did not exist substrates with NANOG 7-12 nucleotides upstream of the center of the octamer in the bottom enrichment, whereas there are 7 such substrates in the top enrichment. Refer to Table S3 for the calculation of test statistic.

Additional figure in pdf format:

Figure S3, related to Figure 4: Gibbs Sampling Motifs for Top Enrichments. Refer to Supplementary Experimental Procedures: Gibbs Sampling for information on the presentation of motif logos.

Tables

Top 1% Enriched Set	Mean	Standard Deviation	Times Genomic Average	Enriched Oligos
POU2F1 (formerly known as OCT1)	0.268	0.407	3.71	294
POU5F1 (formerly known as OCT4)	0.271	0.407	3.75	313
NANOG	0.270	0.408	3.74	321
SOX2	0.214	0.386	2.96	360
Anonymous Differentiated	0.211	0.358	2.91	290
Anonymous ES	0.195	0.360	2.70	341
GENOMIC AVERAGE	0.072	0.223	1.0	

Table S1: Phastcons Scoring for Top 1% Enrichment

Experiment	(S1+S2)	L	Area Ratio $\frac{(S1+S2)}{(L+S1+S2)}$
POU2F1 (formerly known as OCT1)	3.47	10.80	.243
POU5F1 (formerly known as OCT4)	16.86	30.44	.356
NANOG	2.53	19.16	.117
SOX2	0	4.38	0
Anonymous Differentiated	17.72	23.54	.429
Anonymous ES	6.24	16.61	.273

Table S2, related to Figure 3: Binding Width Ratio Calculation. Areas were computed by summing the log ratio enrichment value at each nucleotide position in the scheme shown in Figure 3 and Figure S1. See the Supplementary Experimental Procedures: Wide Binding in PluCRs for additional description of the method.

Bin	Top 5% NANOG Enrichment	Bottom 25% NANOG Enrichment
(1) > 7 nt upstream	6 (6.2)	7 (6.8)
(2) 2-7 nt upstream	**7 (3.3)**	**0 (3.66)**
(3) < 2 nt upstream OR downstream of octamer	38 (41.5)	49 (45.5)

Table S3, related to Figure 6: Chi-square Table for Nanog recognition element to Octamer Distance**.** All oligonucleotides with an instance of an octamer motif [ATG(CIT)A(AIT)AT] and an instance of the Nanog recognition element [CAAT] were identified and assigned to bins based on the location of the Nanog recognition element relative to the octamer. We find that the top 5% NANOG enrichment had more substrates in the position 2-7 nucleotides upstream of the octamer motif than the bottom 25% of Nanog enrichment. Using the above table (expected counts in parentheses) in a chi-square test for equality of distributions shows significance with $p = 0.016$. Refer to Figure S2 and Supplementary Experimental Procedures.

Four additional tables:

Table S4: Log Ratio Scores for All Oligonucleotides

Table S5, related to Figure 2: MORE and PORE Oligonucleotide Scores

Table S6, related to Figure 4: TRANSFAC/JASPAR Centroids

Table S7, related to Figure 7: Enriched 5-mer Words in POU2F1 (formerly known as OCT1) Top 1% Enrichment (Dr. Octamer)

Experimental Procedures

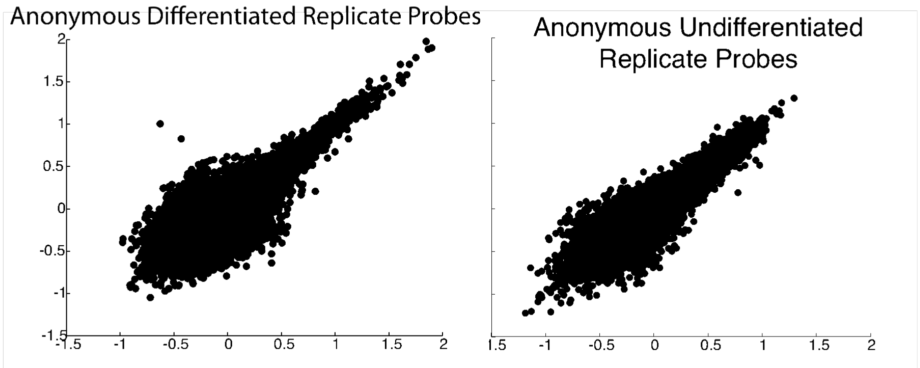
Microarray Analysis

Microarray scores for all 45,973 oligonucleotides were normalized by dividing each probe intensity by the summed intensity of the entire array. Then a log ratio of the normalized red (Cy5 labeled) intensity value to the normalized green (Cy3 labeled) intensity was calculated as the “enrichment score.” During the enrichment scoring, probes with green channel scores lower than the background threshold were removed from the experiment dataset. In about 70% of cases, at least one of the replicate probes for each oligo was included.

To generate a UCSC Browser (Kent et al. 2002) WIG-formatted representation of the data, we computed an average score for each nucleotide position in the hg17 genome. Each genomic coordinate of the 316 PluCR regions received the score that is the mean of all oligonucleotides overlapping that position. UCSC Genome Browser tracks are available from http://fairbrother.biomed.brown.edu/data/pluripotent-2010.

Correlation Between Array Replicates (log red/green ratios):

Experiment	r	r^2
anondiff	.8723	0.76090729
SOX2	.6642	0.44116164
POU5F1 (formerly known as OCT4)	.8117	0.65885689
POU2F1 (formerly known as OCT1)	.6136	0.37650496
NANOG	.9341	0.87254281
anones	.9192	0.84492864



Wide Binding in PluCRs, related to Figure 3

To characterize wide areas of binding in the PluCRs, we computed an average plot of transcription factor binding adjacent to the top enriched portion of the experiments. The top 1% enriched oligonucleotides were located, and a 200bp scoring window was constructed by scanning 100 nucleotides in both directions from the center of the oligo. The nucleotide scores at each position in the window were aligned and averaged, creating a plot like that below for the POU5F1 (formerly known as OCT4) experiment.

To demonstrate wide binding beyond that of a single oligonucleotide, we compared the average binding in the regions greater than 15bp from the center of the enriched oligonucleotide. The central enriched ligand is 30bp long and thus the observed binding outside of the area covered by this ligand suggests clusters of binding sites in the PluCRs. We quantified the extended binding by taking the area between the microarray enrichment curve and a background value. The background value was found by calculating the average enrichment at a sufficient distance from the enriched oligo (i.e. where the curve of log enrichment ratios “flattens” far away from the ligand *L*). A ratio of the area in the regions -100 to -16bp (*S1*) and 16bp to 100bp (*S2*) to the total area above background over the entire 200bp window (the enriched ligand and the above-background regions: $L + S1 + S2$) was computed and recorded as a relative score of binding width. Table S2 gives the associated values for each transcription factor and the anonymous assays. For the example of POU5F1 (formerly known as OCT4) visualized in Figure S1, the area ratio is computed as follows:

$$\frac{(S1+S2)}{(L+S1+S2)} = \frac{16.86}{16.86+30.44} = .356$$

Phastcons Scoring

For each of the six microarray experiments (POU2F1, POU5F1, NANOG, SOX2, anonymous differentiated, and anonymous undifferentiated), a top 1% enrichment of oligonucleotides was determined. Using the UCSC Table Browser, all Phastcons records having any overlap with the enriched oligos in each experiment were used to generate summary statistics. For comparison to a background score, a genome-wide average of the Phastcons records is also provided. (Siepel et al. 2005) See Table S1.

POU2F1 (OCT1) Predictive Words (Dr. Octamer algorithm)

The sequences of the top 1% enrichment in POU2F1 (formerly known as OCT1) were extracted and split into all possible substrings of length 5. From this population of 5-mer words, we computed the count of all 1,024 5-mer nucleotide sequences. Using these data, we estimated the probability of occurrence of the 5-mer words in the POU2F1 (OCT1) enriched oligonucleotides and compared them to the background rate over all 316 PluCR regions.

To assess the statistical signficance of the 5-mer words, we calculated a chi-square test statistic. The statistic test is computed from the following contingency table, where the first quantity is the observed value and the value in parentheses is the expected value. Suppose that *A* is the enriched set of interest in which we wish to prove the significance of a k-mer word. Let the k-mer word have *a* occurrences amongst all the words in *A*, which we notate as */A/*. All PluCR regions define a background set of words *B*, from which *b* of the occurrences are the same k-mer word as that for which we seek to prove enrichment in *A*.

Chi-square Table for Dr. Octamer

	enrichment	**background**
k-mer word	$a \ \ (b\frac{\lvert A\rvert}{\lvert B\rvert})$	$b \ \ (b)$
all other words	$\lvert A\rvert - a \ \ (\lvert A\rvert - b\frac{\lvert A\rvert}{\lvert B\rvert})$	$\lvert B\rvert - b \ \ (\lvert B\rvert - b)$

Computing the sum of the squared difference between observed and expected for all cells produces a chi-square test statistic with one degree of freedom. Note that the predominant term in the chi-square value measures the difference between the enrichment and the expected number of that k-mer word in the PluCR regions.

A power calculation concluded, given the background frequencies in the PluCRs, that a two-fold difference (between *"enrichment"* and *"background"*) in the k-mer word could be detected for 96% of all possible 5-mers. 5-mers were chosen for further analysis on the basis that only 50% of 6-mers would show statistical significance under the scenario of two-fold enrichment.

Scores for all 1024 5-mer words were computed from the set of top 1% POU2F1-enriched oligonucleotides, from which we selected only those with greater occurrences than the background rate ($a > b\frac{|A|}{|B|}$). Table S7 lists all of the enriched 5-mer words with their corresponding chi-square values.

The Dr. Octamer algorithm scanned all 100-mer windows of the *bona fid*e mus musculus POU5F1 regions found by ChIP in Loh et al. 2006. Within each 100-mer window, all 5-mer substrings are matched to the chi-square value of both the forward and reverse complement sequence and averaged. (Words with rates lower than background were scored as 0 for the purposes of this average.) As an additional selection procedure, each of the 100-mer windows in mouse were mapped to conserved regions in other mammalian species (homo sapiens, rattus norvegicus, bos taurus, and canis familiaris). These genomic windows were scored using the same algorithm. We produced UCSC Genome Browser tracks to visualize both the Mus musculus results and the mammalian results (mapped to the mm8 genome build). From the over one thousand POU5F1 regions identified in Loh et al. 2006, six regions were chosen for ChIPseq assay. Three of these regions had exceptionally high predicted POU2F1 binding as well as high predictions in the mammalian species; the other three had very low predicted POU2F1 binding over the entire length of the region as well as low predictions in the mammalian species.

MOREs and POREs

Palindromic POU (OCT) recognition elements (PORE) and more PORE (MORE) were identified in the PluCR regions by scanning each position of those regions for a match to a known binding model. POREs matched either the forward motif "ATTTGNNAKGCAAAT" or its reverse complement "ATTTGCMTNNCAAAT" (Botquin et al. 1998; Remenyi et al. 2001). MOREs matched one of the models "ATGCN(1-3)ATGCAT", the reverse complement "ATGCATN(1-3)GCAT", or the MORE+2 sequence "ATGCATNNATGCAT" identified by Tomilin et al. 2000.

MOREs and POREs within the PluCRs are listed in Table S5, where each row gives the POU2F1 (formerly known as OCT1) and POU5F1 (formerly known as OCT4) enrichment for each of the oligonucleotides containing the element. Due to the tiling scheme, more than one oligo may contain the element in its entirety. Additional sheets are given in the S5 spreadsheet with tolerances for up to two base mismatches in the model sequence.

Figure 3A displays the average percentile enrichment of all substrates which matched the perfect octamer site (ATGCAAAT or its reverse complement), all substrates which match a MORE model, and all substrates that match a PORE model. To make the probability of matching the MORE and PORE approximately equal at any one position in random sequence, we accepted up to one mismatch in MORE and up to two mismatches in the PORE element.

Gibbs Sampling

For each of the six microarray experiments (POU2F1, POU5F1, NANOG, SOX2, anonymous differentiated, and anonymous undifferentiated), a top 1% enrichment of oligonucleotides was input into the Gibbs Centroid Motif Sampler. (Thompson W.A. et al. 2007; Thompson et al. 2003) The Gibbs Centroid Motif Sampler samples from thousands of alignments of the oligonucleotides and returns the centroid motif model. We ran the Gibbs sampler with an increasing number of expected 10-mer motifs in the input model, ascending from 1 to 5. The output of Gibbs sampling is visualized in the form of Berkeley weblogos (Crooks et al. 2004) on an additional supplemental document. Gibbs motif finding was applied to each microarray experiment as well as to the POU2F1 specific enriched portion (with POU2F1 in the top 1%, and POU5F1 microarray score < 0), the POU5F1 specific enriched portion (POU5F1 microarray score > 0.66, POU2F1 < 0), and the intersection of the POU2F1 and POU5F1 top 1% enrichment.

In general, the Gibbs Centroid Motif Sampler assembles more sites when using a simpler motif model. The single motif model may be interpreted as a consensus sequence. As the number of expected motifs increases, the sampler infers more nuanced motifs which have fewer sites in the input sequences.

Supplemental Figure S3 displays all of the image logos for the motifs found in the experiments and special sets described above. For each page, the motifs are printed in order of single motif model (top) to 5-motif model (bottom).

TRANSFAC/JASPAR Motif Scanning

We performed a comprehensive transcription factor binding site (TFBS) search over the 316 PluCR regions. The TFBS prediction was performed with the MotifScanner module of the TOUCAN suite for bioinformatics, (Aerts et al. 2003; Aerts, Van Loo et al. 2005) using all position weight matrices (PWMs) in the TRANSFAC 7.0 Vertebrate database (Matys et al. 2003) and JASPAR CORE database (Bryne et al. 2008). In all, 651 position weight matrices were tested. A Bayesian prior parameter of 0.02 was used to threshold the predictions to high scoring matches.

Using a perl script, the motif predictions in the PluCRs were mapped to all of the oligonucleotides which contained the motif entirely. From these mappings, we calculated a “centroid” score equal to the mean oligonucleotide microarray enrichment score. The POU2F1 and POU5F1 TFBS centroids were then visualized on the scatterplot Figure 4B. In the generation of the scatterplot, all TFBS mapping to a single oligonucleotide were removed, and all non-vertebrate TFBS in the JASPAR database were omitted. The POU2F1 and POU5F1 centroid values are reported for all assayed position weight matrices in Table S6.

Multiple binding models were scored in cases where the TRANSFAC Vertebrate database contained several matrices for the identical factor. The TFBS most enriched in POU2F1 and POU5F1 are POU2F1 binding models. The most enriched non-POU2F1 model was FOXO1 (labeled 5 in the upper-right rectangle of Figure 4B). The FOXO1 motif-bearing oligonucleotides were located and *RAB5A* and *ERMN* gene were targeted as good candidates for competition assay on the basis of having multiple FOXO motifs and an octamer recognition element.

The TRANSFAC and JASPAR motifs found in the PluCR regions at the threshold defined by the prior parameter 0.02 are posted as UCSC Browser (Kent et al 2002) BED format files at http://fairbrother.biomed.brown.edu/data/pluripotent-2010.

Nanog-to-Octamer Distance

All oligos with one mismatch or less from the perfect octamer sequence (ATGCAAAT/ATTTGCAT) were selected from the PluCR regions. To determine an optimal NANOG-to-Octamer distance, oligos from this set with the Nanog recognition element CATT were considered for further analysis. Among these oligonucleotides selected for CATT and octamers, the top 5% enriched NANOG substrates were compared to the bottom 25% enrichment in NANOG to show a difference in distribution.

The top 5% enrichment set had several occurrences of NANOG-to-Octamer distances between two and seven nucleotides upstream of the ATGC/GCAT side of the octamer sequence, whereas the bottom 25% NANOG enrichment was deficient in this region. See Table S3. To calculate a chi-squared test statistic, we binned each distribution into: 1) the positions upstream of this region of difference, 2) the region of difference, and 3) the positions downstream of this region of difference. A chi-square test of significance for this 3x2 contingency table gives a p-value of 0.016, with 93% of the chi-square test statistic contributed by region (2).

PluCR Mapping to Chicken

The PluCR regions were first scored by Dr. Octamer on the basis of their human sequence. This information was used to separate gene regions into quartiles based on their predicted affinity for POU2F1 (formerly known as OCT1) binding. We hypothesized that those regulatory regions with strong POU2F1 binding would have higher conservation in egg-laying vertebrates (those with POU2F1 and not POU5F1 while regions lower in POU2F1 binding would be more likely to undergo divergent mutations.

To test the hypothesis we ran UCSC liftOver (Schwartz et al. 2003) on the coordinates of the PluCR regions and the coding exons for the nearby genes (genes within 5000bp downstream of a PluCR) with parameter minMatch = 0.1. We found whether the regulatory PluCR successfully mapped to a site in chicken and whether at least 99% of the length of coding exons successfully mapped to the correct gene in chicken. From this, we partitioned the PluCRs into P/P (both regulation and gene present), A/P (regulation absent, but gene present), and A/A (both regulation and gene absent). For each set, there are fewer than five P/A's, which were interpreted as liftOver false positives. Within each set, we scored a retention ratio as shown in Figure 7.

References

Aerts, S., G. Thijs, B. Coessens, M. Staes, Y. Moreau and B. De Moor. 2003. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucl. Acids Res., **31(6)**: p. 1753-1764.

Aerts, S., P. Van Loo, G. Thijs, H. Mayer, R. de Martin, Y. Moreau and B. De Moor, 2005. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucl. Acids Res., **33**(suppl_2): p. W393-396.

Botquin, V., V. Botquin, H. Hess, G. Fuhrmann, C. Anastassiadis, M.K. Gross, G. Vriend and H.R. Schöler. 1998. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes & Development. **12**(13): p. 2073-2090.

Bryne, J.C., E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard and A. Sandelin. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucl. Acids Res., **36**(suppl_1): p. D102-106.

Crooks, G.E., G. Hon, J.M. Chandonia and S.E. Brenner. 2004. WebLogo: A Sequence Logo Generator. Genome Research, **14**(6): p. 1188-1190.

Loh, Y.-H., Q. Wu, J.L. Chew, V.B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K.Y. Wong, K.W. Sung, C.W. Lee, X.D. Zhao, K.P. Chiu, L. Lipovich, V.A. Kuznetsov, P. Robson, L.W. Stanton, C.L. Wei, Y. Ruan, B. Lim and H.H. Ng. 2006. The POU5F1 (Oct4) and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet, **38**(4): p. 431-440.

Matys, V., E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender. 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucl. Acids Res., **31**(1): p. 374-378.

Reményi, A., A. Tomilin, E. Pohl, K. Lins, A. Philippsen, R. Reinbold, H.R. Schöler and M. Wilmanns. Differential Dimer Activities of the Transcription Factor Oct-1 by DNA-Induced Interface Swapping. Molecular Cell, 2001. **8**(3): p. 569-580.

Schwartz, S., W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler and W. Miller. 2003. Human–Mouse Alignments with BLASTZ. Genome Research, **13**(1): p. 103-107.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W. J. Kent, W. Miller and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, **15**(8): p. 1034-1050.

Thompson, W., E.C. Rouchka, and C.E. Lawrence, 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucl. Acids Res., **31**(13): p. 3580-3585.

Thompson, W.A., L.A. Newberg, S. Conlan, L.A. McCue and C.E. Lawrence. 2007. The Gibbs Centroid Sampler. Nucl. Acids Res., **35**(suppl_2): p. W232-237.