Positional distribution of human transcription factor binding sites

Supplementary Information

1. ChIP-chip data

This work is primarily based on ChIP-chip data taken from 2 studies performed in Richard Young’s lab at the Whitehead Institute. One study aimed at mapping the binding sites of 3 transcription factors (TFs) known to play central roles in the maintenance of key properties of embryonic stem cells [1]. The second study concentrated on 6 TFs believed to be critical for the biology of hepatocytes which comprise the bulk of the liver [2]. Both studies used human cells and the same custom platform developed in Young’s lab. For 6 of the 9 TFs there is data from 2 biological replicas; for the remaining 3 – HNF6, USF1, CREB1 data from only single replicas were available.

1.1 Brief description of the 9 TFs

1.1.1 Embryonic stem cells related TFs: NANOG, SOX2 and OCT4

Named after the mythological Celtic land of the ever young "Tir nan Og" , NANOG is believed to be a key factor in maintaining the main properties of embryonic stem cells - pluripotency and self renewal [3, 4]. Together with OCT4 and SOX2 it forms a tightly interconnected transcriptional circuit. It was shown that this circuit can act as a bistable switch [5] which can be the mechanism selecting between maintenance of the stem cell phenotype and differentiation. According to [2] these 3 TFs bind to many common target genes. OCT4 and SOX2 often bind to DNA as a heterodimer.

1.1.2 TFs tested in hepatocytes

The 6 transcription factors HNF1A, HNF4A, HNF6, FOXA2, USF1 and CREB1 are known to be expressed in hepatocytes [2].

The transcriptional regulators HNF1A, HNF4A and HNF6 are required for normal function of liver and pancreatic islets. Mutations in HNF1A and HNF4A are the causes of the type 3 and type 1 forms of maturity-onset diabetes of the young (MODY3 and MODY1) [6]

The genes encoding FOXA (hepatocyte nuclear factor 3) family of proteins play a pivotal role in the regulation of metabolism and in the differentiation of metabolic tissues such as the pancreas and liver. FOXA transcription factors bind to cis-regulatory elements in hundreds of genes encoding gluconeogenic and glycolytic enzymes, serum proteins and hormones. Genetic analysis in mice has shown that FOXA2 is necessary for the development of the foregut endoderm, from which the liver and pancreas arise [7].

The ubiquitously expressed USF-1 and USF-2 proteins interact with high affinity to cognate E-box regulatory elements (CANNTG) which are particularly represented over the genome. The USF transcription factors are key regulatory elements of the transcriptional machinery mediating recruitment of chromatin remodelling enzymes, interacting with co-activators and members of the pre-initiation complex (PIC). USF transcription factors have been involved as key regulators of a wide number of gene regulation networks including stress and immune response, cell cycle and proliferation [8].

The cyclic AMP (cAMP)-responsive element-binding protein (CREB) is a ubiquitously expressed transcription factor that is involved in transcriptional regulation of many different cellular processes, it stimulates the expression of numerous genes in response to growth factors, hormones, neurotransmitters, ion fluxes, and stress signals [9].

Table S1: This table lists the DNA binding domain types of the 9 TFs and the HGNC [10] approved names of the corresponding genes.

TF name as appears in this work	**The HGNC name of the coding gene**	**DNA binding domain family**
NANOG	NANOG	homeobox
OCT4	POU5F1	homeobox
SOX2	SOX2	HMG box
HNF1A	TCF1	homeobox
HNF4A	HNF4A	nuclear hormone receptor
HNF6	ONECUT	CUT
FOXA2	FOXA2	fork-head
CREB	CREB1	basic-leucine zipper (bZIP)
USF1	USF1	basic helix-loop-helix (bHLH)

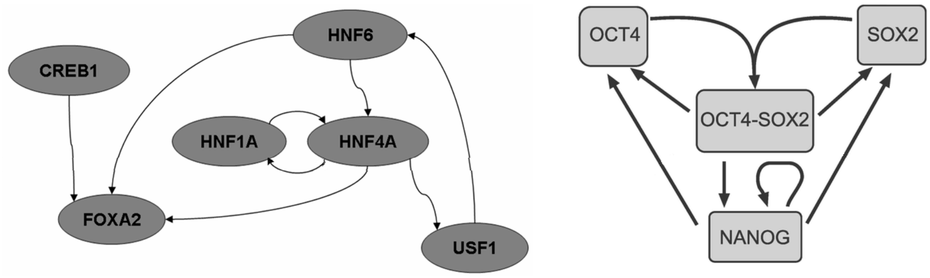


Figure S1: Circuits of the liver (left) and embryonic stem cell (right) related genes as reported by ChIP-chip studies. Arrows indicate regulation but not necessarily activation (it may be suppression). The stem cell circuit diagram on the right was adopted from [5].

1.2 Essential details of the experimental platform

This section describes minimal details of the experimental platform that are essential for understanding the description of data analysis that will follow. A full account of the technique can be found in supplementary material of [1] and [2] or on the web site accompanying those publications [11].

The technique is based on custom designed DNA microarrays containing 60-mer oligonucleotide probes. The probes are covering regions from 8kb upstream to 2kb downstream from transcription start sites of about 18,000 annotated human genes. The average probe density inside the covered regions is approximately one probe every 280bp. This microarray was code-named 10array because it actually consists of 10 separate slides due to limitation of number of probes that can be fitted on a single slide.

After immobilizing the proteins and fragmenting the DNA (into fragments of length of 550 bps on average) part of the resulting material is used for immunoprecipitation, while the other part is reserved for control. The immunoprecipitation enriched DNA extract is labeled with red fluorescent dye while the control whole cell DNA extract is labeled with a green fluorescent dye. The whole cell extract is assumed to contain any piece of the genome at equal probability (concentration) as opposed to the immunoprecipitated DNA extract that contains significantly increased concentrations of DNA fragments to which the TF of interest was bound. Both DNA extracts are applied to the microarray to allow competitive hybridization. The fluorescence intensity is then measured in red and green filters separately for each probe.

1.3 Data analysis pipeline

A probe that has a binding site located within several hundred base pairs from it is expected to give high reading in the immunoprecipitated (IP) channel as compared to whole cell extract (WCE) channel. The probe intensity in the IP channel is a function of (among other effects) the binding site strength, its distance from the probe and the probe's affinity to its matching DNA (which depends on probe's sequence). The latter appears to contribute the most to the probe-to-probe variation but is expected to be canceled out by comparing the IP and WCE channels.

1.3.1 Normalization of raw data

The array set consists of 10 slides due to limitation of the number of probes that could be printed on a single slide. A detailed description of the array design can be found in the supplementary material of [1]. The design files and the raw data were downloaded from [12]. The raw data is in GPR format a description of which can be found at [13].

The purpose of this normalization procedure is to eliminate the effects of variation in experimental parameters (such as the amount of hybridized DNA) between the slides and between the red and green channels. The normalization procedure used in this work follows closely the procedure describe in supplementary material of [1] with slight variations.

Apart from the probes covering promoter regions, each slide contains the same set of control probes. There are negative controls designed not to bind any human DNA, as well as intensity controls that, based on test hybridizations, give signal intensities that cover the entire dynamic range of the array.

The normalization procedure is as follows:

```
1. Filter out all the probes with signal to noise level (as defined in GPR file) greater than 2
2. For each channel (red and green) :
   2.1. For each slide:
        2.1.1. Calculate the median of intensity controls
        2.1.2. Divide all the readings by this median
   2.2. Join the readings from all slides into a single list
   2.3. Calculate the average of the 10 medians calculated in step 2.1.1
   2.4. Multiply all the readings by this median.
   2.5. Calculate the median of negative controls and subtract it from all readings
3. Scale all the readings in green channel so that the medians of all experimental probes in both channels will be equal.
```

1.3.2 M-scores

In order to compare the intensities of the red (IP) and green (WCE) channels, the raw readings normalized using a procedure described above. Log ratios (base *e*) of the red and green channels were calculated as *LogRatio*=log(red)-log(green). However, as can be seen on Figure S2, the differences between log intensities in the two channels diminish with increasing average intensity and therefore, a single fold-change cutoff cannot be used to determine which probes report binding sites.

To cope with this problem, a special score was developed which we called the M-score in analogy with the X-score used for a similar purpose by the authors of [1] and [2]. Roughly, the M-score of a probe is its log ratio divided by the standard deviation of log ratios of all the probes on the array that have "similar" average intensity, where average intensity is defined as log(red) + log(green).

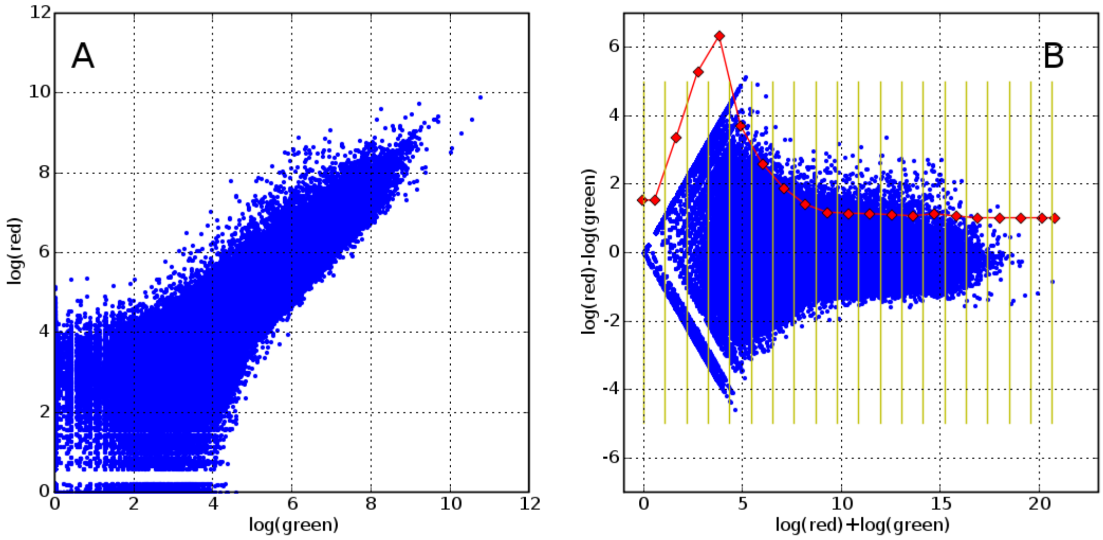


Figure S2: A) Scatter plot of log intensities of the red and green channels B) This graph illustrates the calculation of M-scores.

In more detail, the M-score is calculated as follows: For each probe, the log ratio is calculated as log(red)–log(green) and the average intensity as log(red)+log(green). The range from 0 to max average intensity is split into 20 equal intervals shown by yellow lines on Figure S2 B. Standard deviation of log ratios is calculated in each interval and the resulting number is assigned to the middle of the interval. The red line on Figure S2 shows the local standard deviation times 3. The local standard deviation for each probe is then calculated as linear interpolation between the mid points of intervals. M-score of a probe is its log ratio divided by the local standard deviation. This method was developed in analogy with [14].

The resulting score has unit variance and approximately normal distribution. Therefore M-score cutoffs can be interpreted in terms of probability. That is, the probability of getting a score m_0 or higher is $P(\text{M-score}>m_0) = 1\text{-NormCDF}(m_0)$ or, alternatively, for a given p-value of 0.001 the M-score cutoff would be 3.09.

1.3.3 Bound probes and regions

As most probes were spaced within the resolution limit of chromatin immunoprecipitation, triplets of closely spaced consecutive probes were required to provide evidence of a binding event in order to filter out single probes with spurious signal. Average M-scores for triplets of consecutive probes spaced at less than 1000bp of each other were calculated as $M^{(3)}=\left(M_{left}+M_{mid}+M_{right}\right)/\sqrt{3}$. Under the assumption of statistical independence of M_i and $M_{i\pm1}$, the smoothed variable $M_i^{(3)}$ is also approximately normally distributed with unit variance (see table S2 and Figure S3 for verification of statistical independence).

A triplet was labeled as *bound* if it passed the following criteria based on 4 cutoff parameters : t_1, t_2, t_3 and t_n.

- $M^{(3)} > \theta \cdot t_3$
- AND either:
    - $M_{mid} > \theta \cdot t_2$ AND ($M_{left} > \theta \cdot t_2$ OR $M_{right} > \theta \cdot t_2$)
    - OR
    - $M_{mid} > \theta \cdot t_1$ AND ($M_{left} > \theta \cdot t_\text{n}$ OR $M_{right} > \theta \cdot t_\text{n}$)

The two latter criteria seem redundant, but the cutoffs were selected so that they would cover two different situations, one where a binding event occurs midway between two probes and each detects the event, and the other where a binding event occurs very close to the central probe and is very weakly detected by a neighboring probe.

The θ parameter was introduced in order to study the effect of varying the cutoffs with a single parameter as described in section 2.3.

These cutoffs were adopted from [2] as follows:

- $\text{SF}(t_3)=0.0001$
- $\text{SF}(t_1)=0.0001$
- $\text{SF}(t_2)=0.0005$
- $\text{SF}(t_\text{n})=0.05$

Where SF(x)=1-NormCDF(x) is the survival function of the normal distribution. Central probes of triplets that passed those filters were marked as *bound probes*.

Table S2: Correlation coefficient between M-scores of consecutive probes, standard deviations of single M-scores and of M-score triplets.

TF	Correlation(M_i , M_{i+1})	Std(M)	Std(M_3)
FOXA2_M1	0.16	1.05	1.20
FOXA2_M2	0.08	0.99	1.06
NANOG_M2	0.19	1.01	1.17
NANOG_M3	0.21	1.00	1.17
OCT4_M1	0.19	1.00	1.14
OCT4_M4	0.11	1.00	1.10
HNF4A_M1	0.26	1.05	1.27
HNF4A_M2	0.16	1.03	1.16
HNF1A_M1	0.18	1.12	1.29
HNF1A_M2	0.11	1.07	1.16
USF1_M1	0.13	1.04	1.16
SOX2_M3	0.20	1.27	1.49
SOX2_M4	0.19	1.05	1.22
CREB_M1	0.16	0.97	1.09
HNF6_M1	0.24	1.11	1.33

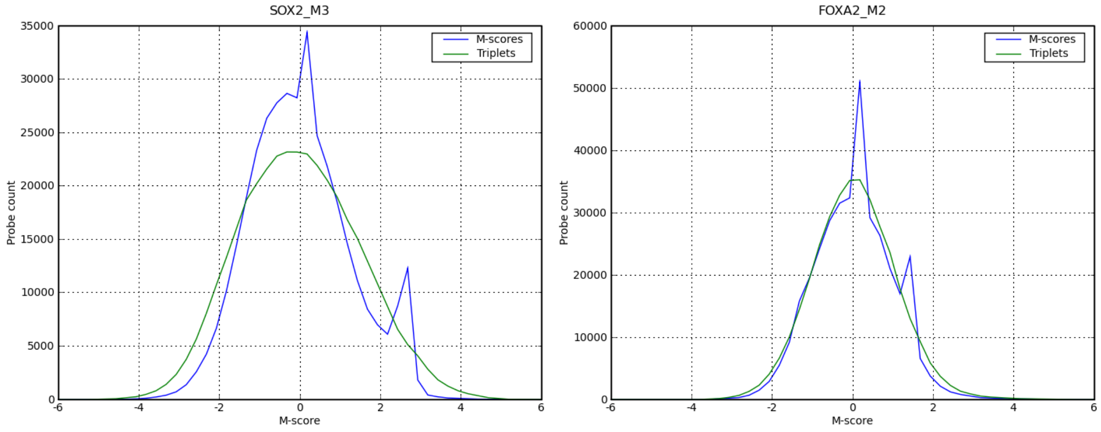


Figure S3: Typical distributions of M-scores for single probes (blue) and for triplets (green).

1.3.3.1 Bound regions

For each triplet that passed the filters, the region between the two of its flanking probes was marked as *bound region*. Overlapping regions were collapsed into single region. For example, on Figure 1 (in the main text) there are two consecutive probes detected as bound (labeled with red triangles), the region between them extended up to the nearest unbound probes is marked as bound region (labeled with magenta line on the figure).

The bound regions were originally intended to be fed into motif discovery algorithms such as [15] or [16] but later proved to be useful for the analysis described below.

Table S3: Example of several actual regions detected as bound from HNF1A data. All the genomic addresses refer to UCSC hg17 genome build.

Chromosome	Start	End	Length
chr6	161093061	161093901	840
chr5	132236487	132237603	1116
chr1	95250176	95250941	765
chr5	35083572	35084350	778
chr4	155839325	155840026	701
chr1	74908903	74909636	733
chr6	31731168	31732010	842
chr1	70588042	70589141	1099
chr1	158006268	158007477	1209

2. Positional distribution of transcription factor binding sites

2.1 Alignment of bound regions relative to TSS

The preprocessing steps described in the previous chapter resulted in lists of several hundred to several thousand bound regions for each TF. Each bound region is several hundred bps long (700 on average). In order to estimate the distribution of binding sites as a function of distance from transcription start site (TSS), the bound regions were aligned relative to the nearest TSS and a number, which we called "coverage number", was calculated. It is somewhat similar to a histogram, but since the bound regions all have different lengths, a simple histogram could not be used. The coverage number of a nucleotide at a given distance from the TSS is the number of bound regions for which this point, i.e. the nucleotide is within the bound region. That is, we count how many bound regions *cover* a point at distance *x* from the TSS, adding up for all the genes tested. Figure 2 in the main text illustrates this concept. The genomic locations of the genes were taken from RefSeq genes table from UCSC genome browser, build hg17.

Put more mathematically, let *R* be the set of all regions detected as bound, we define a_r and b_r as the distances of the two endpoints of a bound region *r* from the TSS nearest to this bound region. As a convention $a_r < b_r$. The distance is defined to be negative if the point is upstream of the TSS and positive if it is downstream, that is, inside the gene. Then the coverage number $Cn(x)$ is given as:

$$Cn(x) = \sum_{r \in R} I(a_r < x < b_r) \qquad (5.1.1)$$

Where *I*(condition) is an indicator function giving 1 if the condition is true and 0 otherwise. Example plot of the coverage number can be seen on Figure S4.

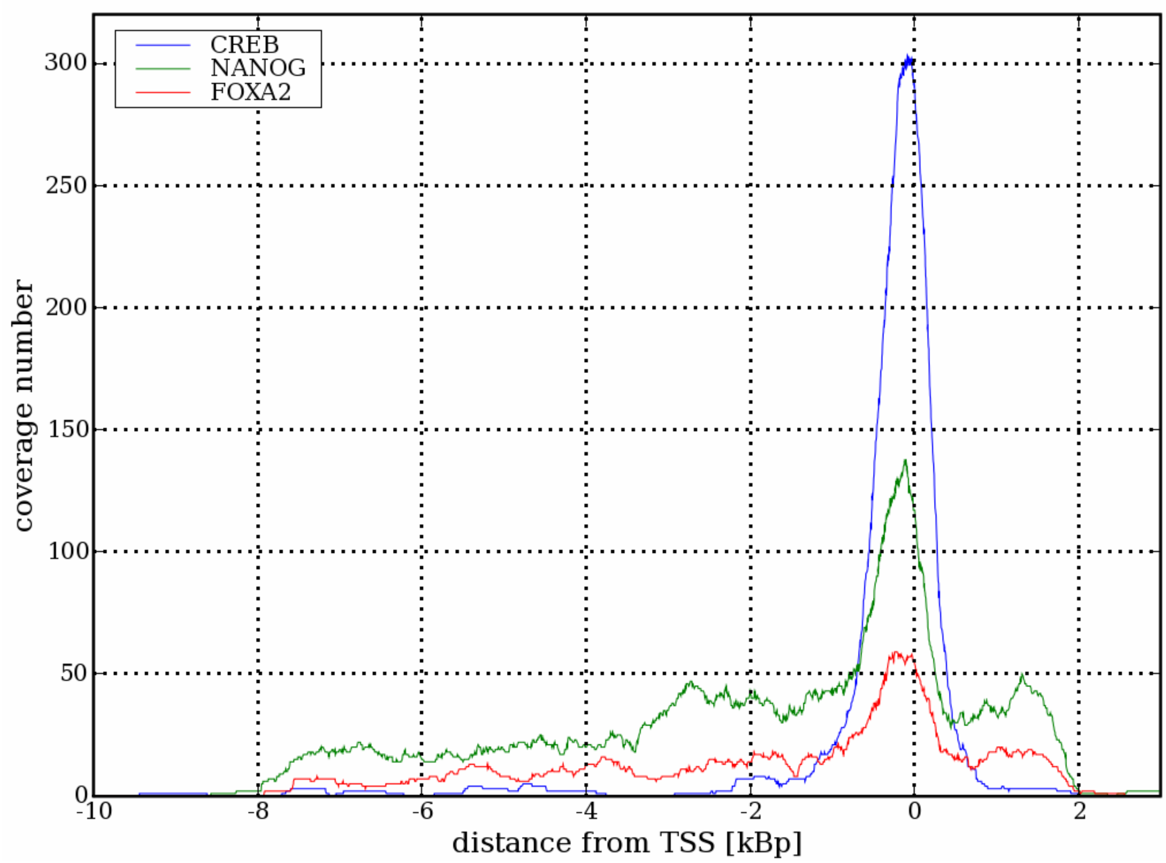
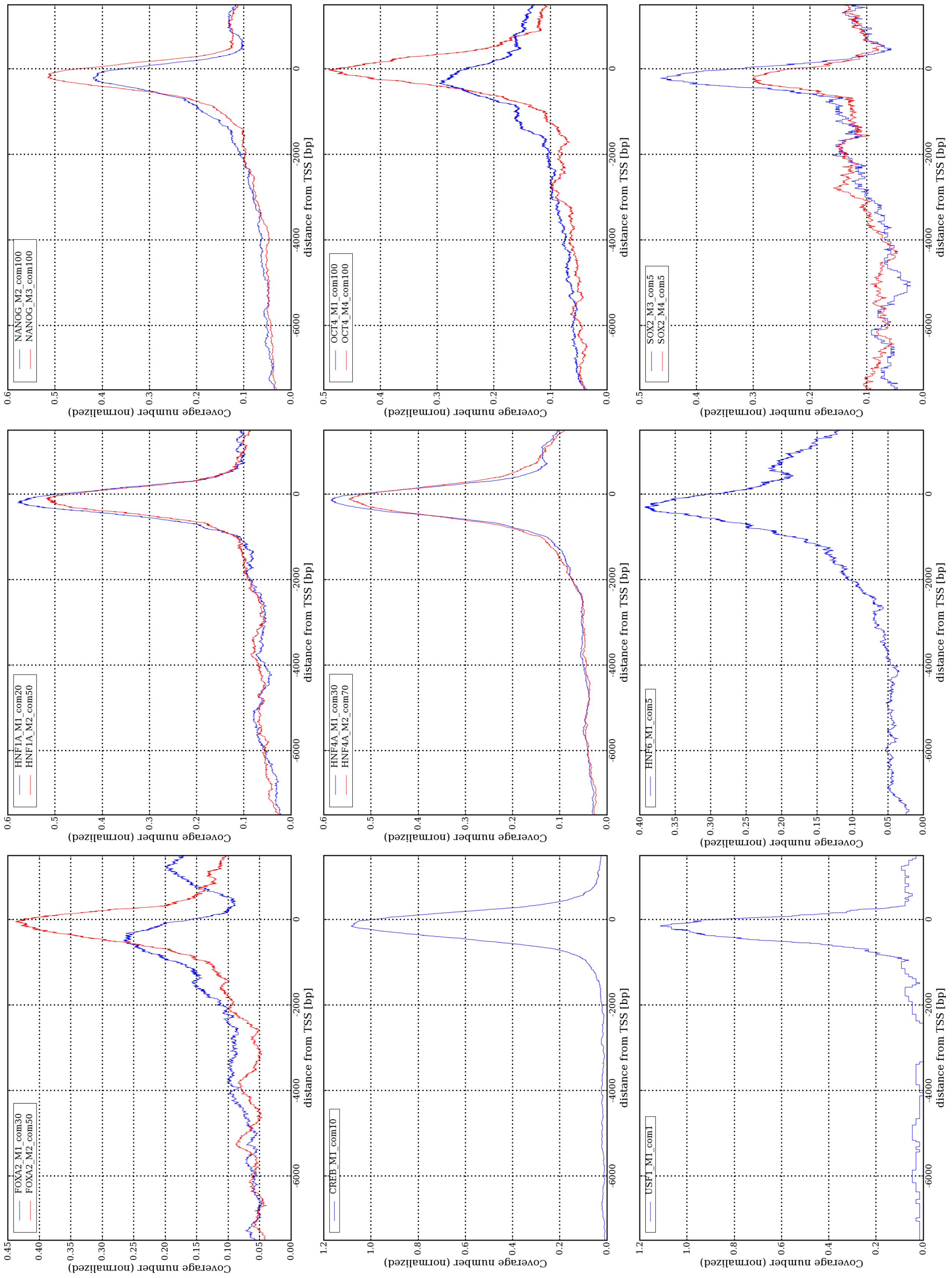


Figure S4: An example of coverage number plot. *x* axis shows the distance from TSS, negative distance means upstream and positive downstream (i.e. inside the gene).

Figure S5 (Over the page): The raw coverage number plots of all the TFs including biological replicas (replicas are labeled as M1, M2 …). There is no correction for probe density variation on these plots. Section 2.4 will introduce a procedure which will result in graphs with probe density taken into account. The score cutoffs shown as "com#" in the legends were selected as described later in section 2.3.


FOXA2_M1_com30
FOXA2_M2_com50
HNF1A_M1_com20
HNF1A_M2_com50
NANOG_M2_com100
NANOG_M3_com100
CREB_M1_com10
HNF4A_M1_com30
HNF4A_M2_com70
OCT4_M1_com100
OCT4_M4_com100
USF1_M1_com1
HNF6_M1_com5
SOX2_M3_com5
SOX2_M4_com5
Coverage number (normalized)
distance from TSS [bp]
-6000
-4000
-2000
0

Looking at the 6 TFs with duplicates on Figure S5, it can be seen that reproducibility can be quite poor. For two TFs – FOXA2 and OCT4 the duplicates are very different from each other.

The most prominent feature on all the plots is the sharp peak just upstream of the TSS. For USF1 and CREB the vast majority of detected binding sites are concentrated within this peak. Other TFs seem to have, in addition, a considerable number of bindings sites almost evenly spread over all the distances

Thinking of binding site localization statistics a priori – before seeing the plots on Figure S5, one would hypothesize that binding site distributions would be roughly bell shaped curves with different widths and positions of the maxima for different TFs, maybe some distributions would even be bimodal. But instead, this data suggests that the binding site distributions of the 9 tested TFs look like a mixture of two distributions - a uniform one and a sharp peak. What vary between different TFs are not the constituent distributions but rather the weights of their contributions to the mixture.

The above picture of a mixed distribution suggests that there might be two distinct groups of binding sites which differ in their biological function or in the mechanism by which their function is achieved. An attempt to find such a difference is presented in section 2.5.

The lower part of the graph for NANOG and FOXA2 seems to decrease gradually away from the TSS and fall to nearly zero at -8kb and +2kb. This effect is due to microarray design. The microarray does not cover promoter regions outside -8kb to 2kb from the TSS, which explains the zero count outside this region. Additionally, the probe density within the covered regions is not uniform. As can be seen on Figure S6 (red curve) the probes are placed more densely near TSS.

In order to understand how this probe density variation influences the coverage number, we performed a simulation of the measurement process starting with a hypothetical TF having a uniform distribution of binding sites as a function of distance from the TSS. Section 2.4 describes the details of how this simulation was performed. The blue curve on Figure S6 is the average of the coverage number plots obtained from 100 such simulations.

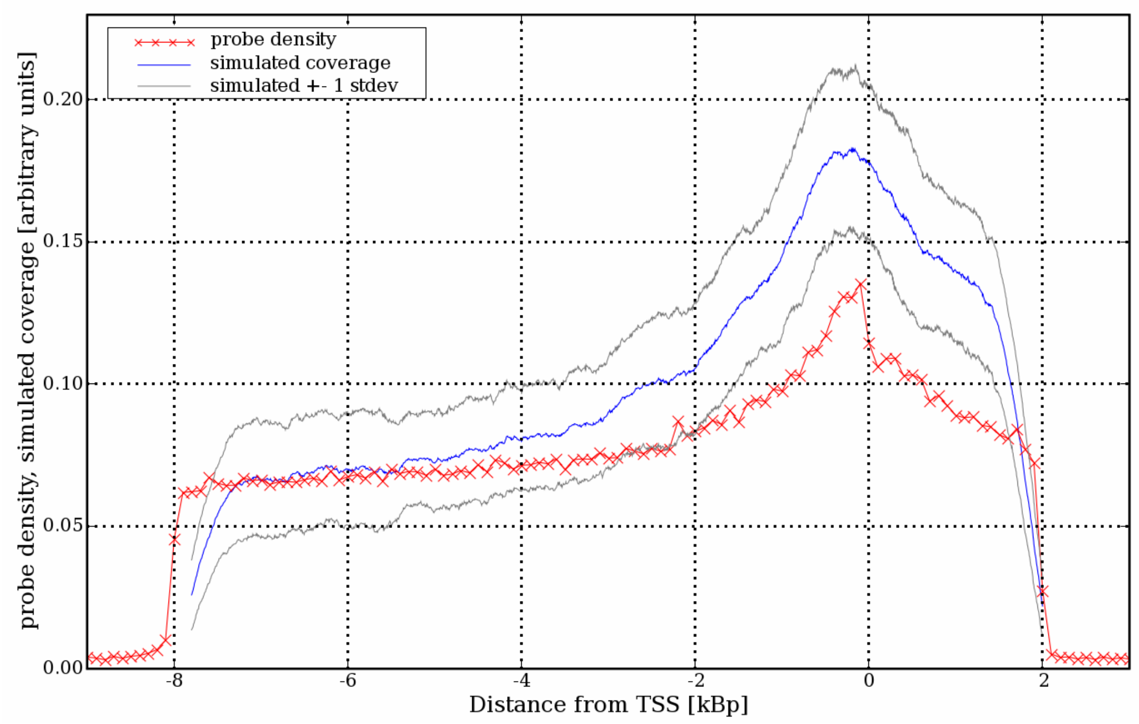


Figure S6: Probe density and the average ± one standard deviation of 100 simulated coverage plots of hypothetical TFs with uniform distribution of binding sites relative to TSS.

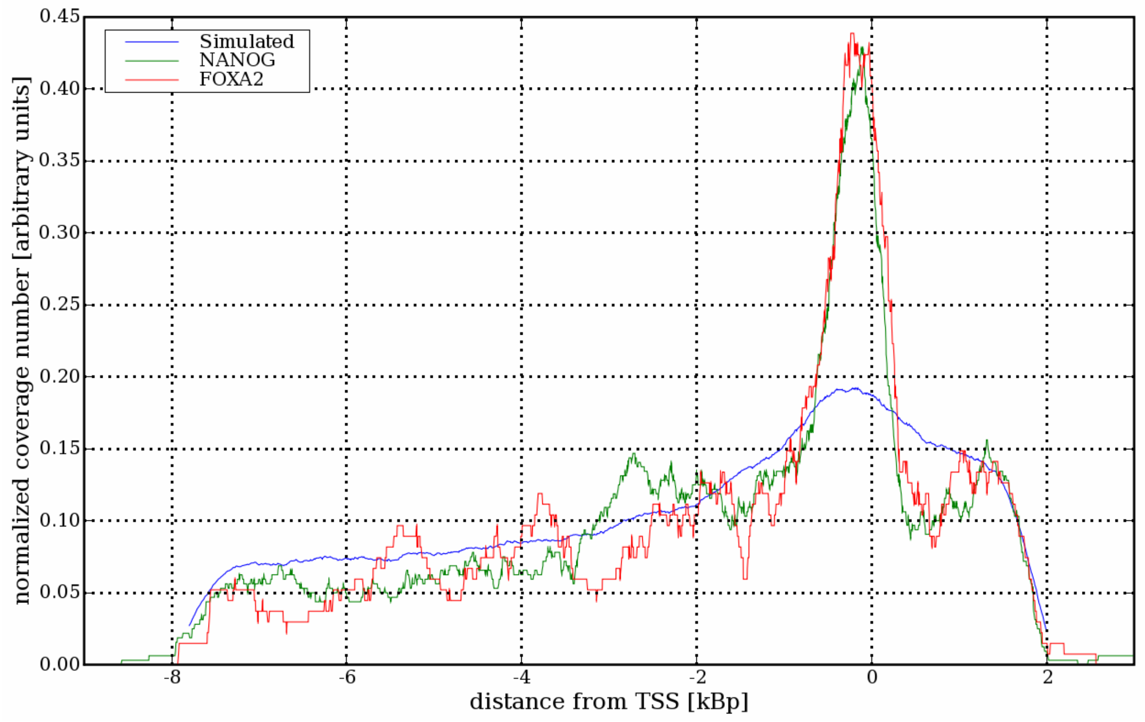


Figure S7: Comparison of coverage number plots for 2 real TFs with simulated coverage number for uniform distribution of binding sites. All 3 curves were normalized to have the same area under the curve, this normalization allows easier comparison between different TFs.

Comparison of this simulated curve with those generated from a real TF data as presented on Figure S7, shows that the gradual decrease towards the edges and the sharper decrease starting at about 400bp from the edges can be explained by differences in probe density. The sharper decrease at the edges is a kind of edge effect. Since this edge effect has great influence on the impression an observer gets from viewing the graphs (even a well informed observer), most of the further graphs will be shown with the *x* axis limits set to conceal the edges in order to avoid false impressions.

Another change that was introduced in Figure S7 is normalization: the coverage number plots are scaled, for each TF and for the simulation, to have the same area under the curve. This normalization allows easier comparison between different TFs. For example, curves for NANOG and FOXA2 that looked different on Figure S5 are now seen to be very similar apart from the total number of bound genes. The area under the graph is roughly the number of bound regions times the average length of a bound region; increasing any of these two parameters would result in higher coverage numbers for the same positional distribution. With such normalization, a higher peak near the TSS means that a greater *percentage* of all the detected binding sites of this TF are located within the peak.

2.2 Comparison with distribution of computationally derived transcription factor binding sites

In order to verify that the findings described above can be reproduced from an independent data source we turned to a computationally derived database of binding sites conserved between human, mouse and rat. This resulted in similar graphs containing the same main features.

The source used is the database underlying the “TFBS Conserved” track in the UCSC genome browser [17, 18]. It was generated using the TRANSFAC [19] collection of positional score matrices (PSSMs) representing the binding preferences of transcriptions factors. The database contains the locations and scores of transcription factor binding sites conserved in the human/mouse/rat alignment. A binding site is considered to be conserved across the alignment if its score meets the threshold score for its binding matrix in all 3 species. The score and threshold are computed with the TRANSFAC Matrix Database (v7.0) created by Biobase [19]. The data are purely computational, and as such not all binding sites listed are biologically functional but the double filter of relatively stringent scores and evolutionary conservation should result in few false positives.

The conserved binding sites were aligned relative to TSS in the same manner as the bound regions from ChIP-chip experiments, but since now the precise location of each binding site is known, a simple histogram can be used instead of the coverage number plots. Figure S8 presents sliding window histograms for HNF1A, USF1, CREB1 and FAC1.

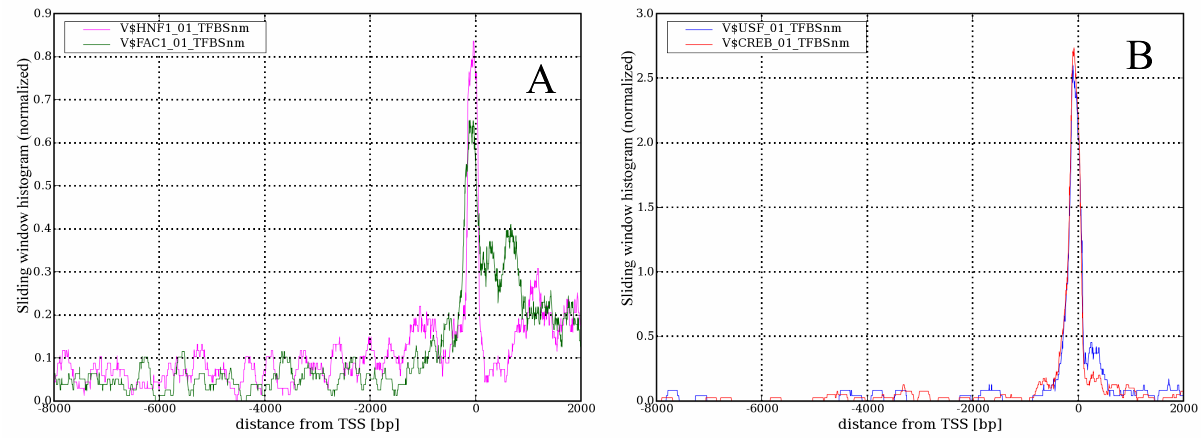


Figure S8: Sliding window histograms of binding site distances from TSS (window size of 200bp). Based on UCSC database of evolutionarily conserved binding sites computationally derived from TRANSFAC PSSMs. It can be seen that the main features are the same as on the coverage number plots based on ChIP-chip experiments on Figure S5.

It can be seen on Figure S8 that the resulting curves maintain the main features of those based on ChIP-chip experiments. Binding sites of USF1 and CREB are, as shown previously, almost entirely concentrated within a narrow peak just upstream the TSS. The peak looks sharper because the sliding window width of 200 base pairs is considerably smaller than the average length of bound regions from ChIP-chip data. The distribution of binding sites of HNF1 has both the peak and an almost uniform component. Unfortunately, most TFs from this database have too few binding sites listed to draw a proper histogram.

2.3 Selection of p-value cutoffs for detection of bound probes

As mentioned in section 1.3 the p-value cutoffs that were used to decide whether a probe is to be counted as bound or not, were rather arbitrary. We decided to explore how changing those cutoffs affects the coverage number plots.

For this end, each of the four p-value cutoffs as given in section 1.3 were multiplied by a single number θ, which we called "cutoff multiplier" (abbreviated as "com" in figure legends) and the whole data analysis pipeline was run with those new cutoffs. The cutoff multiplier was varied from 0.1 to 500 (lower multiplier means stricter cutoff). The coverage number plots for HNF1A and NANOG with a range of cutoff multipliers are shown on Figure 8 (in the main text). Figure S8 shows the number of regions detected as bound as a function of the cutoff multiplier.

Variations of coverage number plots as a function of cutoff could be divided into 3 different types. For 5 TFs the shape remained almost invariant up to some cutoff multiplier, and deteriorated quickly for looser cutoffs until it resembled the simulated coverage plot for a hypothetical TF with uniform distribution of binding sites. HNF1A on Figure 8 (in the main text) belongs to this type. This boundary differs between transcription factors and even between experimental replicas for the same TF. The coverage number plot for HNF4A_M2, for example, remained unchanged all the way up to 100, while HNF4A_M1 started to flatten considerably at 30 as can be seen on Figure S9 (M# indicate experimental replicas).

A different behavior was exhibited by coverage number plots for the stem cell TFs NANOG and OCT4. For them the peak value initially gets higher with increasing cutoff multiplier until it reaches a maximum around com of 100, and then decreases (Figure S9). This behavior may imply different biological roles for binding sites of different affinities, but this hypothesis requires further investigation.

The peaks on the coverage number plots of the remaining 2 TFs USF1 and CREB1 decreased monotonically with cutoff multiplier without apparent discontinuities. It is interesting to note that these 2 TFs have the highest peaks, with coverage numbers nearly zero outside the peak (see Figure S4).

The 1st type of behavior of coverage number plots exhibited by 5 TFs can be used for selecting the best cutoff for each experiment. By starting with a relatively stringent cutoff, one can derive a coverage number plot that corresponds to the distribution of a relatively clean list of binding sites with few false positives. It can be assumed that while the coverage number plot does not change when slightly loosening the cutoffs, the growing list of binding sites maintains a noise level similar to the initial one, but when the cutoff is set too loose, many false binding sites enter the list and the noise level rises affecting the coverage number plot. Therefore, it makes sense to select the loosest cutoff for which the shape of the coverage number plot did not start to deteriorate yet.

In further analysis and on Figure S5 where single cutoff had to be selected per TF this was done in the way described in the previous paragraph (where possible). The cutoffs for

NANOG and OCT4 were selected to get the highest peak. Since the peak heights and number of bound genes for USF1 and CREB1 exhibited no obvious discontinuity (see Figure S8 and S9) and therefore provided no hint for the selection of cutoffs, we selected rather conservative values of *com* (1 for USF1 and 10 for CREB1). This resulted in relatively small numbers of genes detected as bound by USF1 (and to some extent, by CREB1), compared to other TFs and to what was reported by [1, 2].

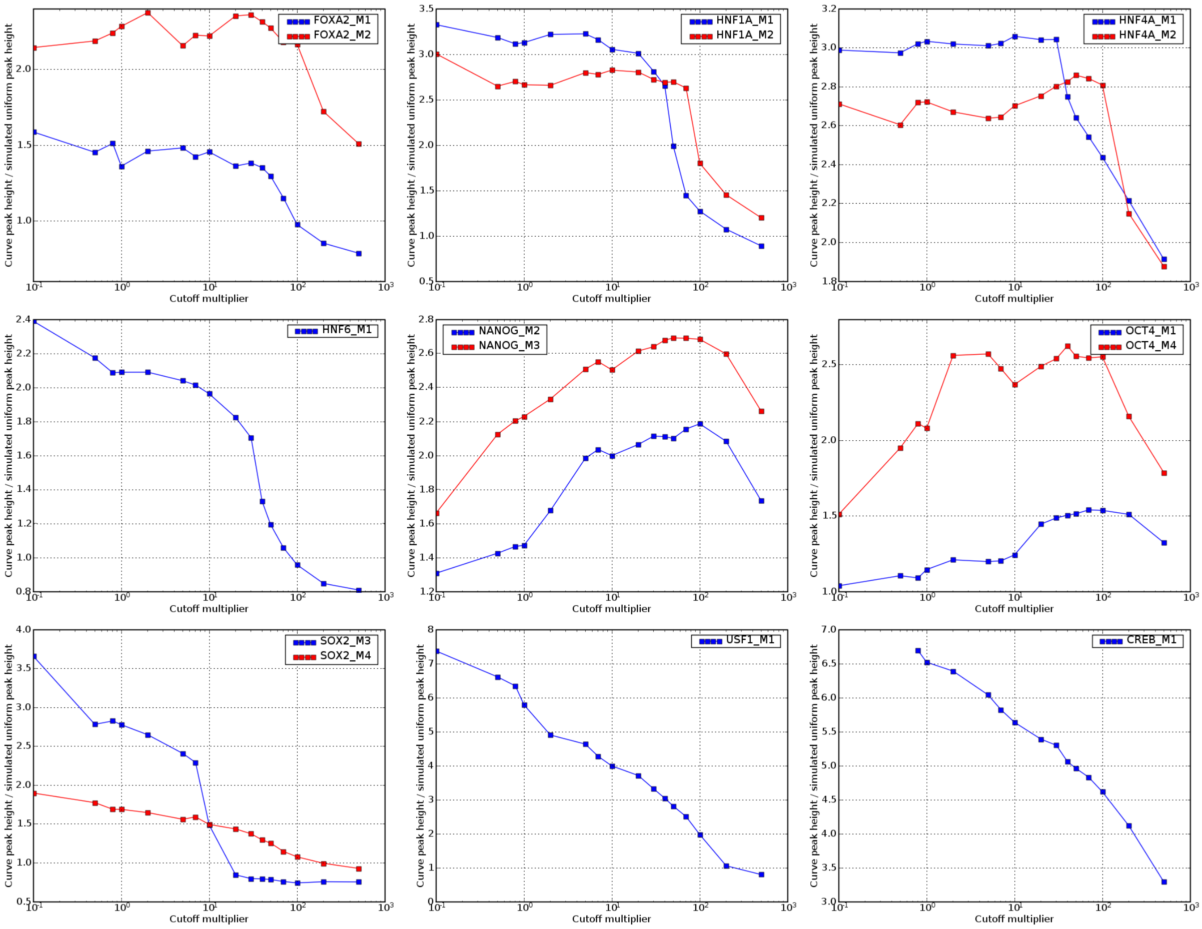


Figure S9: Peak height as a function of cutoff multiplier. The units are such that 1 is the height of the peak for simulation of a TF with uniform distribution of binding sites.

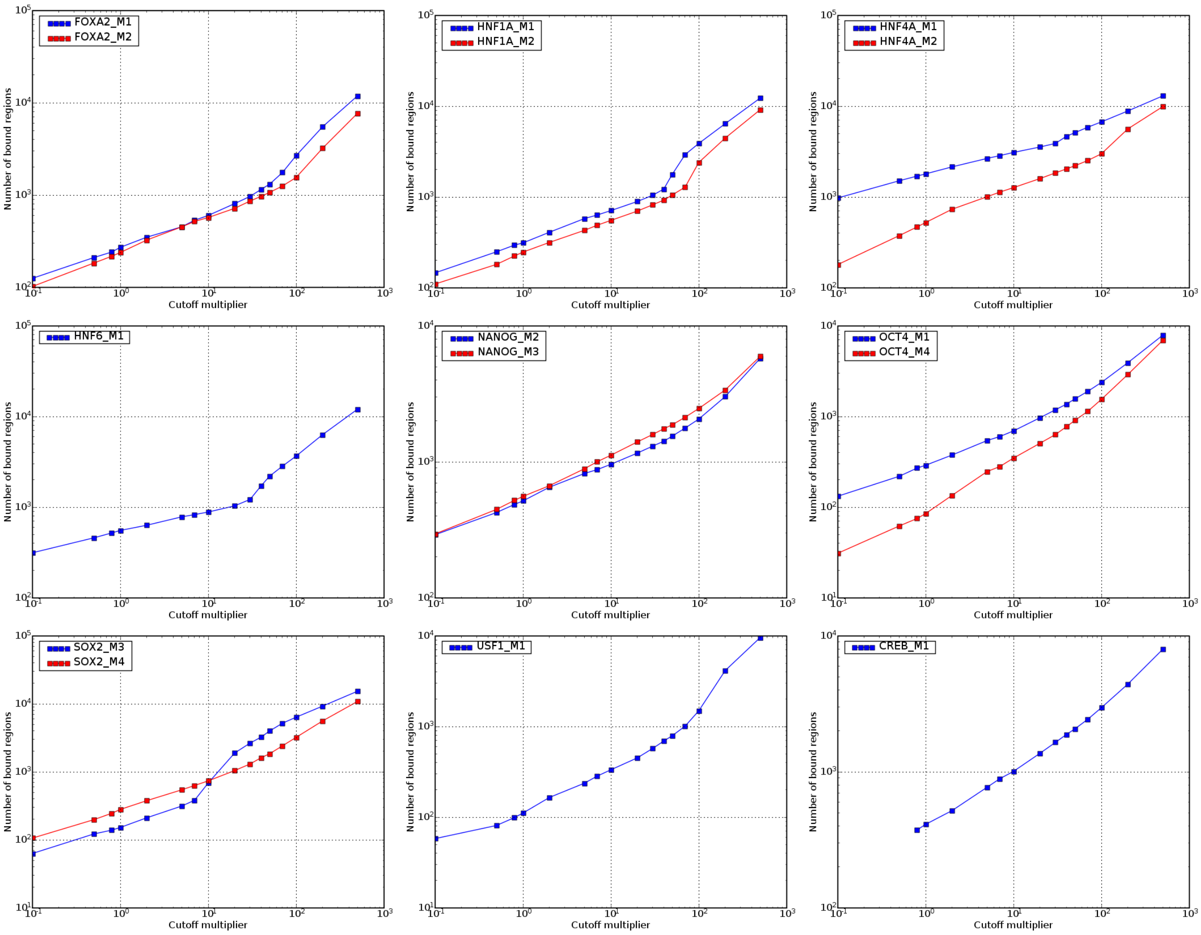


Figure S10: Number of bound regions as a function of cutoff multiplier. Notice that on this plot the boundary for cutoff multiplier is also apparent. M# represent experimental replicas.

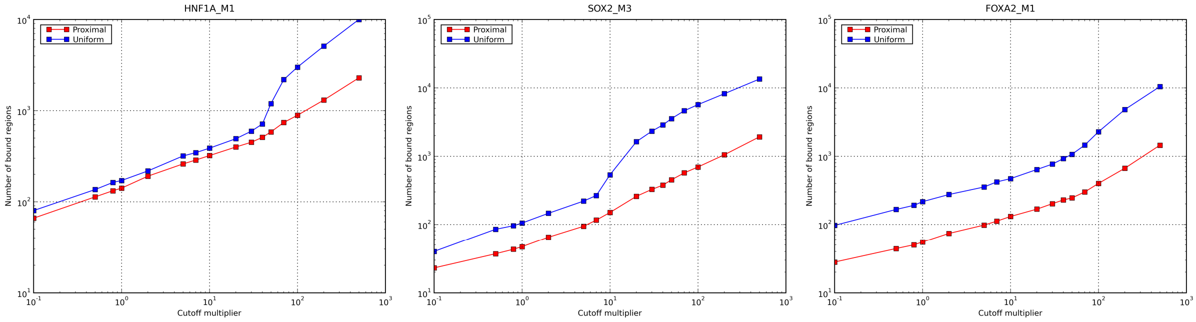


Figure S11: Number of bound regions as a function of cutoff multiplier drawn separately for bound regions that overlap the proximal promoter region (within 300 bp of the TSS) (red) and those that don't (blue). The discontinuity is even more apparent for the distal regions.

2.4 Estimation of distribution of binding site positions

The simulation of the measurement process mentioned in section 2.1 was initially designed to help understand the influence of non uniform probe density and to show how a coverage number plot would look for a hypothetical TF that has uniform distribution of binding sites as a function of distance from TSS. However, with a minimal change, it was extended to handle a fictional TF with any distribution of binding sites. This was used in order to estimate the real distribution of binding sites for each TF.

The simulation accepts a list of microarray probes as genomic addresses and a list of binding sites, each as a genomic address and a strength parameter which plays the role of binding affinity of the site. The results shown below were obtained from simulations with 10,000 binding sites randomly drawn from some distribution (that is tested). For each probe a simulated M-score is calculated as a function of distance d to the nearest binding site and of its strength parameter as given in the equations below. These simulated M-scores are then fed into the analysis pipeline as if they were derived from real raw data and a coverage number plot is generated.

$$\text{M-score} = f(d)\cdot\text{strength} \qquad (1)$$

$$f(d) = \sum_{l=d} l \sum_{a=d}^{l} p(a)p(l-a) \qquad (2)$$

The influence function $f(d)$ used to calculate the M-score was adapted from [20] (the derivation is in the supplementary material), it is based on the distribution of DNA fragment lengths after fragmentation during the ChIP-chip protocol, the longer fragments there are, the more distant binding site can be sensed by a probe, but at larger distances fewer fragments will be long enough to cover the probe.

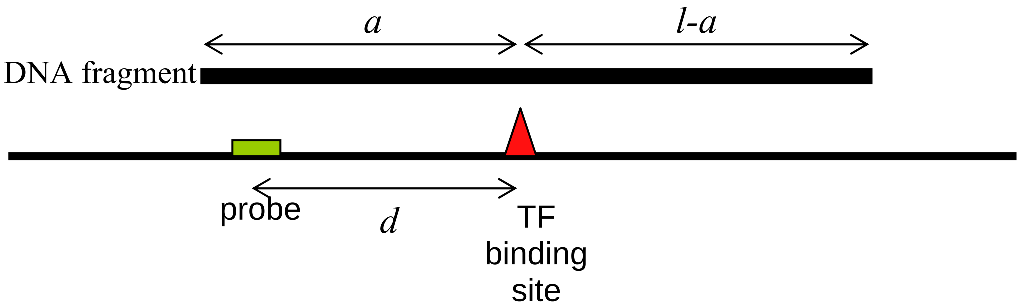


Figure S12: Definition of variables for the calculation of influence function $f(d)$

Following is an outline of the derivation of the influence function $f(d)$ given in eq 5.5.2. Let $p(a)$ be the probability that DNA was cut at a distance a from the TF binding site, then the probability of observing a fragment that was cut at distance a upstream of the binding site and distance b downstream, would be $p(a)p(b)$ since the two cuts are independent. The total probability to observe a DNA fragment of length l would be given by a convolution:

$$p_L(l) = \sum_{a=0}^{l} p(a)p(l-a) \qquad (3)$$

The fluorescent intensity of a DNA fragment increases roughly linearly with its length as more labeled nucleotides are incorporated. Consider a microarray probe with a binding site at distance *d* from it. The total intensity of the probe will be the sum of intensities of all the DNA fragments bound to it which is proportional to a sum over all the possible DNA fragments that contain both – the binding site and the probe, where each DNA fragment contributes in proportion to its length and the probability to observe such a fragment. This sum is given by the equation below, with the substitution *l*=*a*+*b* and change of summation variable this sum can be brought to the form as in equation 5.5.1.

$$f(d) \propto \sum_{a \geq d, b > 0} p(a)p(b)(a+b) \qquad (4)$$

An alternative approach would be to assume that the TF binding site can be found anywhere within the DNA fragment with equal probability. This assumption results in a slightly different influence function *f*(*d*) but the overall result of the simulation remained almost the same.

The total probability to observe a DNA fragment of a particular length $P_L(l)$ was measured experimentally by the authors of [20]. It was approximated with a shifted gamma distribution. A convolution of two identical (shifted) gamma distributions is again a (shifted) gamma distribution with twice the mean (and the shift). Therefore *p*(*a*) was also taken to be a shifted gamma distribution. The following parameters were used for the simulation: shift *s*=50bp, shape parameter *k*=2 and scale parameter *θ*=60.

$$p(x) = \frac{(x-s)^{k-1}}{\theta^k \Gamma(k)} \exp(-\frac{x-s}{\theta}) \qquad (5)$$

The binding strength parameters were randomly sampled from a shifted gamma distribution (shift *s*=3, shape *k*=2 and scale *θ*=3). The gamma distribution was chosen based on the model derived in [21], however the actual distribution of binding strengths should have minimal effect on this simulation.

The list of TF binding sites can be generated with any distribution of distances from the TSS; in particular, we tried to find distributions which, after simulation, would generate coverage number plots very similar to those generated from the real data.

Since any particular coverage plot can be obtained from many different binding site distributions and the simulation is computationally intensive (about 55sec on Intel P4

2.4GHz 1GB RAM for single run), any systematic fitting method would be difficult to implement. The distributions shown below were fitted by hand, and this can be seen as a kind of approximate deconvolution.

In order to verify the manual fitting procedure, we performed a simple grid search in the neighborhood of the manual "best fit" for one of the TFs – USF1. We used the Kullback–Leibler divergence as a measure of difference between the experimental and simulated coverage plot. This resulted in rather similar parameters: peak position, stdev of the peak and the weight of the uniform component were -240, 120 and 1.4 respectively compared to -200, 90, and 1.0 from the manual fitting.

The importance of the deconvolved plots is in their ability to demonstrate that the variation of the coverage plots *outside* the peak can be explained by the non uniform probe density, while the distribution of detected binding sites there is almost uniform.

Figure S13 shows an example of a simulated coverage number plot, together with the distribution from which it was derived, and the real coverage plot for comparison. The fitted distributions for all TFs are shown on Figure 3 (main text). Table 1 (main text) gives a summary of the main parameters – peak position and width, and the relative weight of the uniform component as compared with the localized one.

The fitted distributions support the previously presented hypothesis that the binding sites distribution contains 2 main components: a highly localized peak and a uniform distribution that contribute with different weights for different TFs. In addition, for several of the TFs a minor peak within the gene can be seen. It may correspond to alternative, not yet discovered TSSs or to binding sites that function by directly interfering with RNA polymerase after it started transcribing.

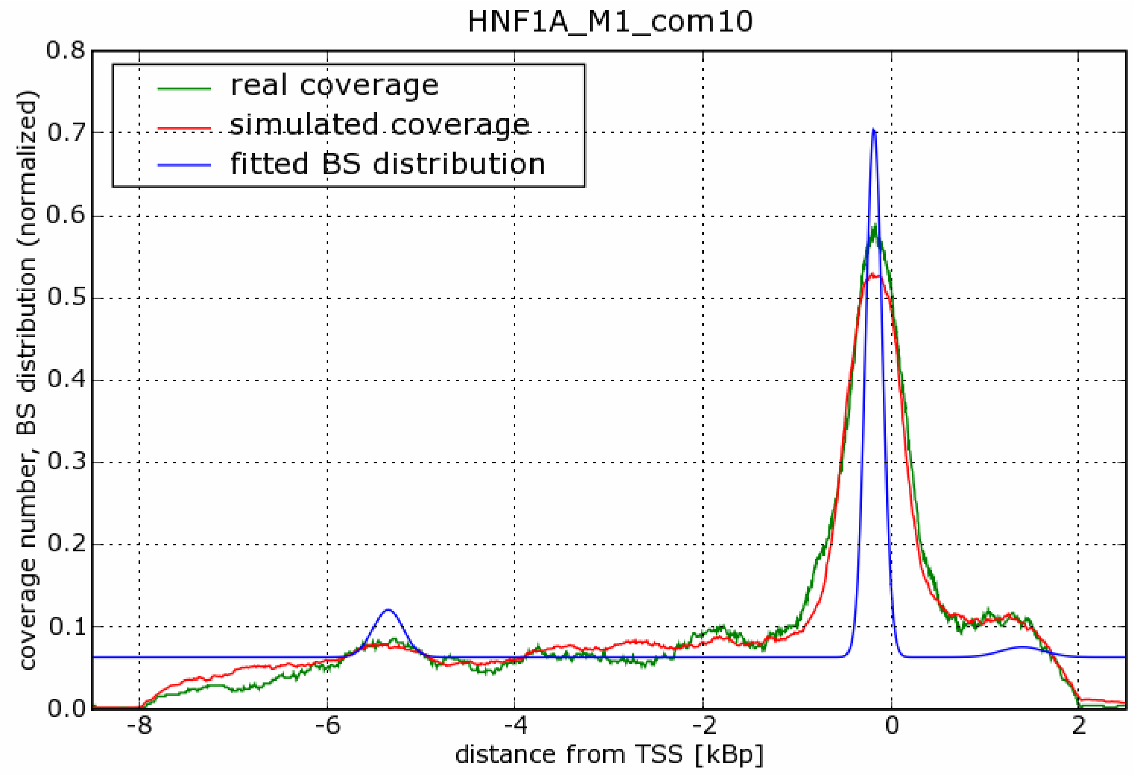


Figure S13: Example of a fitted binding site distribution for HNF1A and comparison of the derived simulated coverage number plot with the real one.

2.5 Comparison of binding sites within the peak and outside using Gene Ontology (GO)

The fact that the binding site distributions can be explained as a sum of a uniform distribution and one with a sharp localized peak suggests that there might be two distinct groups of binding sites which may differ in their biological function or in the mechanism by which their function is achieved. This section presents a simple attempt to find such difference using gene ontology [22] (GO) annotations of genes bound by the 9 TFs according to ChIP-chip data.

For this end the genes bound by a particular TF were split into 2 groups – one contains genes that have a probe detected as bound within 300 base pairs of the TSS and the other contains the rest of the genes with a bound probe within 8kb upstream of the TSS. The first group is assumed to include most of the genes that have a binding site on their promoter within the peak. Both groups were subjected separately to GO enrichment analysis using only the "biological process" type of GO annotations. Results of this analysis are depicted graphically on Figure 4 (in the main text).

The single hypothesis p-values were calculated as typically done in GO analysis using the hypergeometric distribution, separately for each TF, each GO term and each group of genes (far vs. close). Only GO terms containing at least 3 genes from one of the two groups of genes were considered.

For performing the FDR correction (calculating the q-value), all the single test p-values were grouped into the two families according to distance from TSS as described in the previous paragraph. And an FDR correction was performed in each family separately. That is, *N* – the number of tests in a family in each FDR correction was the number of TF-GO pairs.

Figure 4 (main text) shows that for some TFs such as USF1 and CREB1 there are enriched terms only in the group of genes with bound probes close to the TSS. The situation is reversed for OCT4. HNF4A has several GO terms enriched in both the far and the close groups. NANOG has some terms like mitosis enriched only in the close group (see Table 2 (main text) and supplementary table S4), other terms like morphogenesis enriched only in the far, and yet others like RNA metabolism enriched in both (supplementary table S4).

It is interesting to note that for genes bound by stem cell TFs NANOG, OCT4 and SOX2, development related GO categories are enriched only among the genes with a binding site *far* from the TSS (see Table 2 (main text) and supplementary table S4). In contrast, in a group of house keeping genes (derived from [23]) OCT4 and NANOG had much stronger tendency to proximal binding (see Figure S14).

Another observation is that within the circuit of liver related TFs most of the interaction between the TFs are through binding close to the TSS or within the gene.

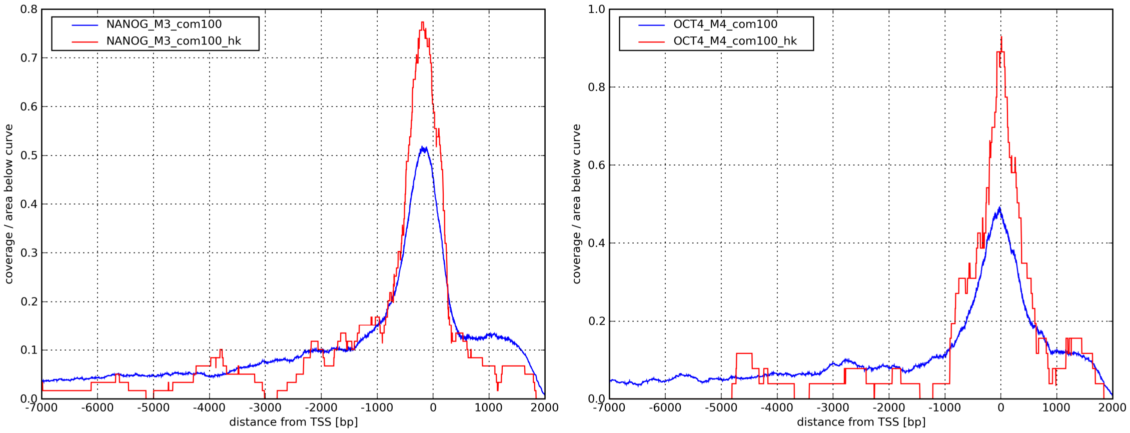


Figure S14: Coverage plots generated from regions bound in the promoters of housekeeping genes (red) compared to those generated from all the bound regions (blue). These plots are in agreement with the general belief that house keeping genes tend to beproximally regulated.

References

1. Boyer, L., et al., *Core transcriptional regulatory circuitry in human embryonic stem cells.* Cell., 2005. **122**(6): p. 947-56.
2. Odom, D., et al., *Core transcriptional regulatory circuitry in human hepatocytes.* Mol Syst Biol, 2006. **2**: p. E1.
3. Mitsui, K., et al., *The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.* Cell, 2003. **113**(5): p. 631-42.
4. Rodda, D.J., et al., *Transcriptional regulation of nanog by OCT4 and SOX2.* J Biol Chem, 2005. **280**(26): p. 24731-7.
5. Chickarmane, V., et al., *Transcriptional Dynamics of the Embryonic Stem Cell Switch.* PLoS Computational Biology, 2006. **2**(9): p. e123.
6. Odom, D., et al., *Control of pancreas and liver gene expression by HNF transcription factors.* Science, 2004. **303**(5662): p. 1378-81.
7. Kaestner, K.H., *The hepatocyte nuclear factor 3 (HNF3 or FOXA) family in metabolism.* Trends Endocrinol Metab, 2000. **11**(7): p. 281-5.
8. Corre, S. and M.-D. Galibert, *USF as a key regulatory element of gene expression.* Medecine sciences : M/S, 2006. **22**(1): p. 62-7.
9. Euskirchen, G., et al., *CREB binds to multiple loci on human chromosome 22.* Mol Cell Biol, 2004. **24**(9): p. 3804-14.
10. Wain, H.M., et al., *Genew: the Human Gene Nomenclature Database, 2004 updates.* Nucleic Acids Res, 2004. **32**(Database issue): p. D255-7.
11. *Accompanying web site for Boyer et. al.* [cited; Available from: http://jura.wi.mit.edu/young_public/hESregulation/Technology.html.
12. *Raw ESC ChIP-chip data from Boyer et al.* [cited; Available from: http://jura.wi.mit.edu/young_public/hESregulation/Data_download.html.
13. *GPR - GenePix Results format (*.gpr).* [cited; Available from: http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html#gpr.
14. Vencio, R.Z. and T. Koide, *HTself: Self-Self Based Statistical Test for Low Replication Microarray Studies.* DNA Res, 2005. **12**(3): p. 211-4.
15. Pavesi, G., et al., *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.* Nucleic acids research, 2004. **32**(Web Server issue): p. 199-203.
16. Siddharthan, R., E.D. Siggia, and E. van Nimwegen, *PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.* PLoS Comput Biol, 2005. **1**(7): p. e67.
17. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
18. Weirauch, M. and B. Raney. *TFBS conserved track at UCSC genome browser.* [cited; Available from: http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=tfbsConsSites.
19. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic acids research, 2003. **31**(1): p. 374-8.
20. Qi, Y., et al., *High-resolution computational models of genome binding events.* Nat Biotech, 2006. **24**(8): p. 963-70.
21. Sengupta, A., M. Djordjevic, and B. Shraiman, *Specificity and robustness in transcription control networks.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(4): p. 2072-7.
22. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
23. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes are compact.* Trends Genet, 2003. **19**(7): p. 362-5.