Supplemental Materials for Mason et al.

Sequence details

Sequences were extended relative to probe or peak width by adding flanking regions to each side. Probes from Sridharan *et al.* were about 60bp wide on average and were extended by 150bp. The median peak widths in Chen *et al.* and Marson *et al.* were 7bp and 224bp, and we extended their sequences by 200bp and 100bp, respectively. In each dataset, overlapping sequences were combined to make a long sequence.

For Oct4 context-dependent motif discovery, OS-cobound sequences were defined as the Oct4 bound sequences that contained Sox2 binding within 50bp (3896 sequences). Oct4-only sequences were those Oct4 bound sequences that lacked Sox2 binding within 5kb (477 sequences). These two distance cutoffs (50bp and 5kb) where chosen to define clearer regions of cobinding and single binding. The same cutoffs were used to define ST-cobound sequences (1749) and Tcf3-only sequences (402).

Oct4 peaks in the Chen study were split into Oct4-motif peaks and pGCAT peaks by scanning for the two motifs. For every candidate segment in an Oct4 bound sequence with LR > 500 for either motif, we designated it a site of the motif that gave a higher LR. Peaks with only Oct4 sites were considered Oct4-motif peaks and those with only pGCAT sites were considered pGCAT peaks. Sequences that had both sites or neither sites were not included in the analysis.

Updating motif length

We illustrate the update of motif length by the calculation on the right side of a motif. The computation on the left side is completely analogous. Label the motif positions as $1, \ldots, w$ and denote by $N_i = (N_{iA}, \ldots, N_{iT})$ the differential nucleotide counts at position i calculated from predicted sites in S_1 and S_2 (eq 3 in the main text). Similarly, we use $N_{(w+1)}$ to denote differential counts on the right flanking positions of the sites. We compute the Bayes factor at position i, for $i = w, w+1$, which is the ratio of the probability that the observed counts N_i are part of the motif (H_a) over the probability that they are from the background Markov model H_0,

$$BF_i = \frac{P(N_i|H_a)}{P(C_{i-1,i}|H_0)}, \tag{1}$$

where $C_{i-1,i}$ denotes the transition counts from position $i-1$ to i, similarly defined by the differential counts between S_1 and S_2. In the above equation, $P(C_{i-1,i}|H_0)$ can be calculated based on the transition probability matrix of the first-order Markov chain for the background model and $P(N_i|H_a)$ can be obtained in closed form assuming a Dirichlet prior for θ_i (the ith row of the PWM Θ). Define $BF_g = BF_{w+1}$ and $BF_s = 1/BF_w$. From eq (1) we see that BF_g and BF_s summarize the preference for growing and shrinking the motif length, respectively, as compared to keeping the current length w. If both BF_g and BF_s are less than 1, we do not change the motif length. Otherwise we grow or shrink the motif length according to the maximum between BF_g and BF_s. If the decision is to grow, we also require that the observed counts at the extended position are significantly different from the expected counts based on the background Markov model ($P < 0.001$) and that the information content of the counts is greater than 0.1, which makes it more conservative to grow motifs.

FDR comparison with simulated control sequences

In Table 4 we computed the false discovery rate of each motif finder with false positives estimated from control sequences, S_2^t. These control sequences were randomly sampled from the genome based on the distribution of the locations of binding peaks. Many of these control sequences lie in promoter regions. As such they may contain actual binding sites of the ChIP TFs though the chances are very small. As an alternative we computed FDRs with simulated control sequences, S_2^{MC}, generated by a second-order Markov Chain with transition probabilities estimated from S_2^t. These simulated sequences contain no TF binding sites. The number of sequences in S_2^{MC} and their lengths were chosen to resemble those of S_2^t with S_1^t remaining unchanged. This false discovery rate, FDR_{sim}, was then estimated by the same way as we did in Section 3.2. Supplemental Table 1 shows the FDR_{sim} of each motif finder with the percent changes relative to CMF's FDR_{sim} in parentheses. As in Table 4 CMF exhibited a lower FDR_{sim} in almost every dataset. In the few cases where CMF was beat the difference was generally small, such as the case of Klf4 in the Chen study where the difference in FDR_{sim} was < 0.02. These results confirm our conclusion on the superior performance of CMF drawn from Table 4 with test control sequences.

Supplemental Table 1. A comparison of motif finding methods with simulated control sequences

	ChIP(Motif)	CMF	DME	FIRE	BioP
Sridharan	Oct4(Oct4)	0.44	0.45 (2)	0.55 (25)	0.49 (10)
	Sox2(Sox2)	0.39	NA	0.55 (41)	0.45 (15)
	cMyc(Ebox)	0.33	0.41 (26)	0.43 (31)	NA
	Klf4(Klf4)	0.25	0.29 (18)	0.37 (51)	0.23 (−7)
	Nanog(Sox2)	0.46	0.61 (33)	0.99 (110)	0.46 (0)
Chen	Oct4(Oct4)	0.30	0.6 (97)	NA	NA
	Sox2(Sox2)	0.18	0.28 (52)	0.5 (170)	0.23 (22)
	cMyc(Ebox)	0.065	0.14 (120)	0.1 (61)	NA
	nMyc(Ebox)	0.13	0.2 (59)	0.19 (49)	NA
	Klf4(Klf4)	0.078	0.12 (58)	0.52 (570)	0.061 (−22)
	Nanog(Nanog)	0.60	NA	NA	NA
	Nanog(Sox2)	0.34	0.47 (38)	0.72 (110)	0.44 (30)
	STAT3(Stat3)	0.18	0.22 (24)	0.46 (160)	NA
	CTCF(Ctcf)	0.17	0.28 (63)	0.54 (220)	0.36 (110)
	Esrrb(Esrrb)	0.13	0.14 (6)	0.33 (140)	0.18 (34)
Marson	Oct4(SoxOct)	0.17	0.28 (64)	0.58 (250)	0.3 (76)
	Sox2(SoxOct)	0.46	0.38 (−17)	0.81 (76)	0.36 (−21)
	Nanog(SoxOct)	0.41	0.54 (33)	0.69 (70)	0.46 (13)
	Tcf3(Sox2)	0.39	0.36 (−7)	0.5 (27)	0.48 (22)

FDRs are presented with the percent increase over the FDR of CMF in parentheses. NA indicates that the method was unable to find the motif.

TF co-occupancy near Oct4 peaks

We calculated the proportions of Oct4/Sox2 cobound peaks and those peaks bound by Oct4 only (Oct4-only peaks) that were co-occupied by the other 10 TFs in the Chen study, similarly to what we did for the Oct4-motif peaks and the pGCAT peaks. The results are reported in Supplemental Table 2. One sees that the Oct4-only peaks are not enriched for cobinding of cMyc, nMyc, E2f1 or Zfx, implied by insignificant p-values for difference of proportions tests. On the other hand, Oct4/Sox2 peaks are enriched for binding by

Nanog, Smad1, Stat3 and Esrrb with very small p-values. However, compared to the result for the Oct4-motif peaks and the pGCAT peaks, the ratio of the proportion of Oct4/Sox2 peaks with cobinding of Nanog, Smad1 or Stat3 over that of Oct4-only peaks tends to be smaller though the p-value is more significant due to a larger sample size. These observations demonstrate the importance of the two Oct4 context-dependent motifs in identifying distinct cobinding patterns of potential co-regulators.

Supplemental Table 2. Proportions of cofactor binding near Oct4/Sox2 peaks and Oct4-only peaks

	Nanog	Smad1	Stat3	Tcfcp2l1	Esrrb
Oct4/Sox2	0.38	0.26	0.54	0.43	0.42
Oct4-only	0.20	0.13	0.44	0.37	0.35
$-\log_{10}(P)$	39	27	7.9	3.3	5.3
	Klf4	cMyc	nMyc	E2f1	Zfx
Oct4/Sox2	0.45	0.52	0.49	0.64	0.47
Oct4-only	0.41	0.54	0.47	0.60	0.46
$-\log_{10}(P)$	2.0	0.6	0.6	1.4	0.1

Tabulated in the same format as Figure 2c.