Supplementary materials

Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships

Yokoyama et al., 2009.

Section S1: Initial parameter estimates

As described in the text, an MPF is modeled as the sum of a background function $C(x)$ and a signal function $H(x)$. The background function is given by

$$C(x) = b + d \bullet c(x) \qquad (1)$$

where $c(x)$ represents the expected frequency of occurrence according to dinucleotide composition, and b and d are free parameters in the model. The signal function $H(x)$ is given by

$$H(x) = \sum_{j=1}^{M} a_j \bullet \exp\left[-\frac{(x-\mu_j)^2}{2\sigma_j^2} \right] \qquad (2)$$

where M represents the number of overrepresentation peaks within the MPF. For MLFs, M can equal either 0 or 1, while M can take on any non-negative integer for MRFs. The parameters a, μ, and σ represent free parameters within the model. Estimating these parameters involves optimizing the log-likelihood of the data using a quasi-Newton method called Broyden's method (1). This method accepts an initial parameter vector θ_0 and iteratively updates the parameters until they converge to a local optimum. As the final model can ultimately be affected by the choice of the initial parameters, MPF estimation is conducted using several different initial parameter values. These initial values are obtained by considering the outlying data points, as their departure from the background frequency can potentially signify positional enrichment. For each of the several initial parameter estimates for a given motif, values for μ are set to the locations of the most extreme outlier points, while the coefficient a is estimated to be the height of the occurrence frequency at this location over the background frequency. The initial σ-value of the Gaussian term is then determined using the occurrence frequency at the surrounding positions. For a single MPF, Broyden's method is conducted separately on each initial parameter vector. By iteratively updating each parameter vector, the method produces different parameter estimates for each initial vector upon convergence. The final parameters are ultimately taken to be those which produce the highest log-likelihood of the data.

We find that parameter estimation is quite robust, as initial parameter vectors at different positions within the location of enrichment typically converge to very similar values. Initial parameter vectors not within the location of enrichment often produce different parameter estimates, although (by definition) they produce lower log-likelihoods and therefore do not affect the final results. The robustness of the parameter estimates is clearly illustrated by comparing the predictions generated during the human and the mouse RefSeq analyses (Table 1 in the text, and Table T2 in Section S4). The location and width of the enrichment was found to be very similar for motifs present in both species, despite the fact that the analysis was conducted upon independent data sets. Moreover, the initial parameter vectors were generated independently during the two analyses; any random differences in the data would therefore affect the initial parameter estimates. Thus, the initial parameter vectors differed for each motif shared between the two species, with different site locations chosen for the initial μ estimate and different initial σ-values. Therefore, the high level of similarity between the predictions attests to the robustness of the methodology.

Initial parameter estimates for MRFs are determined in a similar fashion as those for an MLF. MRF estimation is conducted in an iterative fashion, with the addition of a new Gaussian term at each step (i.e., incrementing the value of M by 1 at each iteration). In the case where the MRF has multiple Gaussian terms, initial parameter estimates are taken to be the final parameter estimates of the previous MRF, along with the additional Gaussian term added to the signal function. The initial values of this (new) Gaussian term are estimated in the same manner as that for the MLF model, as described above. Although this procedure involves estimating several parameters, in practice we have found that the previously determined parameters (i.e., those of the previous Gaussian terms) do not change during subsequent iterations. Thus, each MRF estimation only involves optimizing the parameters for a single additional Gaussian term, and therefore the robustness of the parameter estimates are equivalent to that of an MLF.

Section S2: MPF statistics

The significance of spatial bias for a given motif is determined by considering two different models, a null model and an alternative model. Here, for simplicity, we illustrate this comparison given the MLF model. The statistical model is identical for MRFs, although more than one such comparison may be necessary for MRF estimation.

For any given MLF $g(x)$, the null model comprises only the background function $g(x) = C(x)$. In contrast, the alternative model is given by $g(x) = C(x) + H(x)$, where $H(x)$ incorporates positional bias into the MLF $g(x)$. The complete function model for any MLF is given by

$$g(x) = b + d \bullet c(x) + a \bullet \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{3}$$

(Note Eqs 1 and 2.) Both the null model as well as the alternative model can be defined as in Eq 3, although the null hypothesis is special case of this model where a is set to zero. Thus, the null model is nested within the alternative hypothesis. In general, when nested models meet certain requirements ('regularity conditions'), the 'scaled deviance' of the two models given the data follows a χ^2 distribution (2). Given the parameter vectors θ_{null} and θ_{alt} of the null and alternative models, respectively, the scaled deviance Z is given by

$$Z = 2 \bullet \left[L(D;\theta_{alt}) - L(D;\theta_{null}) \right] \tag{4}$$

where $L(D;\theta_{null})$ and $L(D;\theta_{alt})$ represent the log-likelihoods of the respective models given the data D.

Assuming regularity of the models, $Z \sim \chi^2$ with $|\theta_{alt}| - |\theta_{null}|$ degrees of freedom. More specifically, Z converges asymptotically to the χ^2 distribution assuming a large number of data points (2). However, there are several regularity conditions that must be satisfied in order for Z to converge to a χ^2 distribution. For instance, the 'true' model must exist within the parameter space of the alternative hypothesis. Of course, this is an assumption made whenever one fits a model function to any data. In our case, inspection of the results suggest that our model does accurately fit the data (e.g., note Figure 1a in the text).

Another requirement for regularity is that the null model θ_{null} must be found strictly on the interior of the alternative model. Specifically, any free parameter of θ_{alt} that is fixed within the null hypothesis must lie within a compact neighborhood Θ lying within the parameter space of the alternative model. It is not necessarily obvious that this holds for the MLF model, as the (fixed) parameter a is set to zero in the null model. If a was restricted to be positive, then the null model would lie 'on the boundary' of the alternative hypothesis rather than within the interior (in such a case, a would equal zero for the null hypothesis, and $a > 0$ for the alternative model). However, in our model, a is not restricted to be positive, but can also take on negative values. We note that in such a case, the MLF itself does not take on negative values, as the positive values for $C(x)$ prevent this from occurring (i.e., $C(x)+H(x)>0$ even though $H(x)<0$). Although cases where $a<0$ are extremely rare, this did occur for one of our predicted 6mers (namely, CGCCCC in cluster 9, Table 1 in the text).

In addition to the requirements given above, the log-likelihood of the model as well as its derivatives must exist and be finite at the 'true' values of the parameters. Our model meets this condition; for instance, the maximum likelihood of any single data point is finite. This maximum likelihood is obtained when the value of the MLF equals the observed frequency of the motif at x; thus, the log-likelihood is bounded by this maximum value. We note that this property is dependent upon the fact that the values of an MLF are not normalized across x. It is important to emphasize that the value of $g(x)$ represents, for a particular value of x, the probability of motif occurrence. Thus, $g(x)$ is not a probability distribution $\underline{across\ x}$, as the sum of its values across x does not equal one. This distinction is important; if the values of $g(x)$ were normalized across all positions, the log-likelihood of the function would become unbounded, as the limit of the log-likelihood would approach $+\infty$ as $\sigma \to 0$. This corresponds to fitting the Gaussian term of $H(x)$ to a single data point, at which the height of the Gaussian curve would rise arbitrarily high as σ approaches zero.

Upon implementation, the method uses an F-test to determine the significance of positional enrichment. The statistic F is given by

$$F=\frac{\left[Z(\theta_{alt},\theta_{null})\right]\left[n-\left|\theta_{alt}\right|\right]}{\left[Z(\theta_{Sat},\theta_{alt})\right]\left[\left|\theta_{alt}\right|-\left|\theta_{null}\right|\right]} \qquad (5)$$

where $Z(\theta_j,\theta_k)$ represents the scaled deviance of models j and k, n represents the number of data points, and model Sat represents the 'saturated model'; i.e., the model that optimizes the log-likelihood of the data without limits on the number of parameters. This statistic F follows an F-distribution with $\left|\theta_{alt}\right|-\left|\theta_{null}\right|$ and $n-\left|\theta_{alt}\right|$ degrees of freedom (2); p-values obtained during the analyses described in the text were produced using this statistic.

We note, however, that the scaled deviance converges asymptotically to the χ^2 distribution as the number of data points $n \to \infty$. In order to determine whether fitting the χ^2 distribution is appropriate given the number of data points during our MLF analysis, we conducted a simple test using a completely randomized data set. Namely, we generated a simulated data set assuming a completely uniform distribution of each nucleotide type at each site. We note that this type of simulated data differs substantially from the control sets mentioned in the text (i.e., intergenic sequences and position-specific dinucleotide simulations). That is, in the completely randomized sequences, no higher-order dependencies exist between sites. In contrast, data sets obtained using biological data are not random, as genomic data follows various trends such as high frequencies of repeat sequences, position-specific fluctuations in dinucleotide content, etc.

Conducting a comprehensive 6mer MLF analysis on the randomized data set only produced one prediction with a p-value under 1e-5. This is similar to what we would expect from a randomized data set, as we are testing for multiple hypotheses. Thus, it appears that the number of data points used during MLF estimation is large enough to assume a χ^2 distribution.

Section S3: Putatively novel 7mers of Vardhanabhuti et al

Vardhanabhuti et al (3) reported a list of 168 putatively novel 7mer motif clusters with positional specificity. Their analysis was first conducted by analyzing positional preferences of known protein binding motifs in TRANSFAC (4). They subsequently scanned for spatial bias across the list of 7mers that had been filtered for matches to the known TF binding sites in TRANSFAC. Thus, they presented predictions from this list of (filtered) 7mers as novel motifs. However, inspection of the results showed that a large fraction of these 7mers still matched known binding sites in TRANSFAC. Table T1 shows the 15 top-ranking 7mers reported by Vardhanabhuti et al. The second row shows the ‘novel’ motif clusters predicted during their analysis; the third column shows matches to known cis-regulatory elements in TRANSFAC according to STAMP (5). The corresponding TRANSFAC motif is shown to the right of its binding protein (only aligned sites are shown), along with the STAMP E-value in the last column. The strongest match according to STAMP is shown, and 7mers matching binding sites known to be positionally enriched are also shown below the strongest match. Eleven of the sixteen produced matches to known binding sites at an E-value threshold of 1e-6, while two additional predicted 7mer clusters produced weaker matches (denoted by the (†) character in the 3rd column). Although these last two motifs produced weaker matches, they are identifiable as the CAAT-box (CCATTGG; rank 9) and the YY1 binding site (GAAGATG; rank 12); positional specificity for each was predicted at the same location of enrichment found using the known binding motif sequences.

Rank	Consensus	TF	TFBS	E-value
1	AGATGGC	NF-muE1	AGATGGC	2e-11
2	ATTGGCT	alpha-CP1	ATTGGCT	3e-10
		NFY	ATTGGYT	8e-8
3	CCGACAT	--	--	--
4	CACTTCC	GABP	CnCTTCC	1e-9
		ETS	nACTTCC	1e-8
5	GGTGAGT	--	--	--
6	AGCCAAT	alpha-CP1	AGCCAAT	3e-10
		NFY	ARCCAAT	8e-8
7	GCGGGGC	SP1	GCGGGGn	7e-7
8	GGAAGTG	GABP	GGAAGTG	1e-10
		ETS	GGAAGTn	1e-8
9	CCATTGG	†alpha-CP1	TCATTGG	1e-6
10	GTCAATC	COMP1	GTCAATC	2e-7
11	GACGTAA	CREB	GACGTMW	1e-9
12	GAAGATG	†YY1	nAAnATG	1e-6
13	CTGATTG	NFY	CTGATTG	3e-9
14	TATAAGG	SRF	WATAAGG	1e-7
		†TBP	TATAAAn	7e-6
15	ATGGCGG	E2F	TTSTCGG	3e-7

Table T1

Section S4: Supplementary MLF tables: Mouse data

Table T2

Motif cluster groups exhibiting positional enrichment in mouse RefSeq promoters (6-8). The mouse RefSeq analysis was conducted on a data set of 18,354 non-redundant regulatory sequences in the (-500,+100) range. Matches to spatially biased motifs in humans are shown to the right of the mouse MLF analysis results. The location (μ) and width (σ) of enrichment are given to the right of each motif cluster. The third column shows factor names binding to the known regulatory elements in TRANSFAC (4); putatively novel motifs in mouse are labeled m1-m15. Both the cross-species comparisons and comparisons to the TF binding sites in TRANSFAC were conducted using STAMP (5) (E-value threshold: 1e-6).

Table T3

Positionally enriched motif comparisons between the mouse RefSeq and RIKEN promoter data sets (9,10). The RIKEN promoter data consisted of 1,354 high-quality mouse promoters produced by CAGE-tag data. Many of the positionally enriched motifs found using the RefSeq data could not be detected using the significantly smaller RIKEN data set. We find that this is due to the highly reduced number of motif occurrences at each position within the promoters. As the MLF model was designed to be conducted on a genome-wide level, we find that the method is best applied to large data sets in order to detect motifs with a low overall frequency of occurrence.

Location-specific motif clusters

Mouse RefSeq data						Human RefSeq data		
Rank	p	TF	Consensus	μ	(σ)	Consensus	μ	(σ)
1	7e−135	Sp1(−)	SCYCCKCCCC	−75.7	(51.7)	GCCCCKCCCC	−73.2	(45.0)
2	4e−127	Tbp	ATATAAARGC	−29.9	(1.7)	ATATAAAWR	−29.6	(1.9)
3	3e−116	Tbp	ATATAW	−29.7	(1.8)	TWTATA	−29.9	(2.0)
4	2e−112	Sp1(+)	GRGGGGGGCGKG	−64.6	(42.8)	AGGGGGCGGGG	−65.3	(52.2)
5	3e−103	*m*1	GGTAAG	71.6	(33.9)	CAGGTAAG	74.5	(31.6)
6	2e−102	*m*2	TGTGTGT	−434.8	(212.5)	GTGTGT	−325.6	(234.4)
7	6e−101	Nfy(+)	WCCAATGR	−84.7	(40.1)	AGCCAATCAG	−75.7	(40.7)
8	7e−100	Nfy(−)	SATTGGT	−83.0	(44.1)	CTSATTGGCT	−75.8	(42.7)
9	5e−98	*m*3	CGCCATGGCY	52.4	(34.9)	CRCCATGGA	53.8	(38.0)
10	4e−92	Creb1	GTGACG	−44.5	(34.6)	CGTGACGTC	−47.1	(39.0)
11	3e−90	Elk4(+)	GCCGGAAGTG	−31.5	(37.0)	ACCGGAAGTG	−23.9	(32.3)
12	4e−73	Zfp36I2	RGCGGCG	32.6	(44.4)	CAGCGGCKGC	37.0	(40.7)
13	2e−64	*m*4	GTGAGTG	70.1	(32.7)	GTGAGTG	69.2	(36.4)
14	2e−63	*m*5	WGGTGA	70.1	(34.8)	--	--	--
15	3e−63	Zeb1	AGGTAA	−47.1	(123.8)	--	--	--
16	4e−60	E2f1	TGGCGG	23.8	(16.7)	GATGGCGG	32.9	(22.1)
17	9e−60	*m*6	CGCGCGC	−31.6	(97.1)	GCGCGC	−51.8	(95.0)
18	3e−58	Zeb1	CAGGTA	313.9	(112.4)	--	--	--
19	9e−57	Nfy(+)	GCCAAT	−91.0	(21.9)	AGCCAATCAG	−75.7	(40.7)
20	3e−55	Yy1	GATGGC	26.9	(20.1)	ATGGCC	53.6	(33.9)
21	8e−53	*m*7	CTGCTGCY	55.1	(37.0)	TCTGCTGCT	54.0	(33.5)
22	3e−51	Creb1	CGTCAC	−53.2	(39.5)	TCGTCAC	−47.0	(37.4)
23	7e−50	Yy1	CAAGATGG	16.5	(10.7)	CAAGATGG	23.9	(17.1)
24	3e−49	Nfy(+)	AGCCAA	−96.4	(46.0)	AGCCAATCAG	−75.7	(40.7)
25	7e−49	Nrf1	TGCGCA	−57.8	(46.8)	RTGCGCA	−52.7	(59.8)
26	2e−47	Elk4(−)	CTTCCGG	−15.4	(16.0)	CACTTCCGGT	−21.3	(32.2)
27	2e−44	Zfp219	CCCCCC	−117.6	(99.9)	†CCCACCC	−130.0	(70.2)
28	2e−43	*m*8	CACGCC	−113.3	(90.6)	--	--	--
29	2e−43	Tbp	TAAATAG	−28.8	(1.7)	TAAAAA	−27.8	(0.9)
30	3e−43	Myc	CACGTG	−53.3	(46.1)	CACGTG	−51.0	(50.7)
31	2e−39	Maz(−)	CCCTCC	−61.9	(32.8)	CTCCCTC	−111.0	(100.6)
32	4e−37	*m*9	AAGGTA	149.6	(47.8)	--	--	--
33	9e−36	Zbtb7a	GCCCCC	−66.3	(33.6)	CGCCCC	35.0	(32.2)
34	1e−32	*m*10	ATGGAG	52.5	(36.2)	--	--	--
35	2e−31	Inr	CTCAGTN	−3.0	(0.2)	GCTCAGTCC	−3.0	(0.2)
36	9e−31	Nfy(+)	AATCAG	−74.4	(35.8)	--	--	--
37	5e−30	*m*11	CTCTCT	−350.6	(114.9)	--	--	--
38	2e−25	Inr	TCAGTC	−2.2	(0.5)	CAGTTG	−1.2	(0.5)
39	1e−24	Tead2	CCGCCG	67.7	(32.1)	--	--	--
40	3e−24	Nfy(−)	ATTGGC	−100.0	(16.1)	CTSATTGGCT	−75.8	(42.7)
41	2e−23	*m*12	GCAGCA	28.6	(15.4)	--	--	--
42	9e−23	Nfy(−)	TTGGCT	−70.9	(26.7)	--	--	--
43	3e−21	Inr	GGCAGT	−3.0	(0.2)	CAGTGC	−1.0	(0.2)
44	7e−21	*m*13	GGTGGC	45.1	(29.6)	--	--	--
45	2e−20	*m*14	GGACCC	102.1	(46.4)	GGACCC	78.7	(27.8)
46	9e−20	Inr	CAGTCY	−1.0	(0.2)	GCTCAGTCC	−3.0	(0.2)
47	9e−18	Inr	CACTTC	−1.0	(0.2)	TCACTT	−1.9	(0.5)
48	1e−17	Mef2	AAAATA	202.3	(78.0)	AAAAAT	77.3	(23.1)
49	1e−16	*m*15	GAAGGT	54.4	(38.3)	AAGAAG	96.5	(55.5)

Table T2

Location-specific motif clusters

Mouse RefSeq data						Mouse RIKEN data		
Rank	p	TF	Consensus	μ	(σ)	Consensus	μ	(σ)
1	7e-135	Sp1(-)	SCYCCKCCCC	-75.7	(51.7)	CCCCGCCC	-63.5	(28.0)
2	4e-127	Tbp	ATATAAARGC	-29.9	(1.7)	GNCTATAWAAG	-33.2	(1.3)
3	3e-116	Tbp	ATATAW	-29.7	(1.8)	CTATAT	-31.9	(0.9)
4	2e-112	Sp1(+)	GRGGGGGGCGKG	-64.6	(42.8)	GGGGGCGGGG	-70.2	(29.6)
5	3e-103	*m*1	GGTAAG	71.6	(33.9)	--	--	--
6	2e-102	*m*2	TGTGTGT	-434.8	(212.5)	GTGTGT	-424.1	(27.5)
7	6e-101	Nfy(+)	WCCAATGR	-84.7	(40.1)	CCAATG	-77.9	(0.4)
8	7e-100	Nfy(-)	SATTGGT	-83.0	(44.1)	ATTGGT	-68.1	(0.6)
9	5e-98	*m*3	CGCCATGGCY	52.4	(34.9)	GCCATG	18.1	(9.4)
10	4e-92	Creb1	GTGACG	-44.5	(34.6)	--	--	--
11	3e-90	Elk4(+)	GCCGGAAGTG	-31.5	(37.0)	--	--	--
12	4e-73	Zfp36I2	RGCGGCG	32.6	(44.4)	--	--	--
13	2e-64	*m*4	GTGAGTG	70.1	(32.7)	GGTGAGT	53.1	(22.9)
14	2e-63	*m*5	WGGTGA	70.1	(34.8)	--	--	--
15	3e-63	Zeb1	AGGTAA	-47.1	(123.8)	--	--	--
16	4e-60	E2f1	TGGCGG	23.8	(16.7)	--	--	--
17	9e-60	*m*6	CGCGCGC	-31.6	(97.1)	GCGCGCG	-19.7	(3.4)
18	3e-58	Zeb1	CAGGTA	313.9	(112.4)	--	--	--
19	9e-57	Nfy(+)	GCCAAT	-91.0	(21.9)	CCAATG	-77.9	(0.4)
20	3e-55	Yy1	GATGGC	26.9	(20.1)	AGATGGC	12.3	(2.6)
21	8e-53	*m*7	CTGCTGCY	55.1	(37.0)	--	--	--
22	3e-51	Creb1	CGTCAC	-53.2	(39.5)	--	--	--
23	7e-50	Yy1	CAAGATGG	16.5	(10.7)	CAAGATG	10.0	(2.6)
24	3e-49	Nfy(+)	AGCCAA	-96.4	(46.0)	--	--	--
25	7e-49	Nrf1	TGCGCA	-57.8	(46.8)	--	--	--
26	2e-47	Elk4(-)	CTTCCGG	-15.4	(16.0)	--	--	--
27	2e-44	Zfp219	CCCCCC	-117.6	(99.9)	CCCCCT	-123.1	(70.7)
28	2e-43	*m*8	CACGCC	-113.3	(90.6)	--	--	--
29	2e-43	TBP	TAAATAG	-28.8	(1.7)	AAATAG	-27.1	(0.5)
30	3e-43	Myc	CACGTG	-53.3	(46.1)	--	--	--
31	2e-39	MAZ(-)	CCCTCC	-61.9	(32.8)	CCCTCC	-84.9	(46.1)
32	4e-37	*m*9	AAGGTA	149.6	(47.8)	--	--	--
33	9e-36	Zbtb7a	GCCCCC	-66.3	(33.6)	GCCCCC	-45.0	(3.0)
34	1e-32	*m*10	ATGGAG	52.5	(36.2)	--	--	--
35	2e-31	Inr	CTCAGTN	-3.0	(0.2)	GSCTCAGTGA	-3.5	(0.7)
36	9e-31	Nfy(+)	AATCAG	-74.4	(35.8)	--	--	--
37	5e-30	*m*11	CTCTCT	-350.6	(114.9)	--	--	--
38	2e-25	Inr	TCAGTC	-2.2	(0.5)	TCAGTT	-2.3	(0.5)
39	1e-24	Tead2	CCGCCG	67.7	(32.1)	GCCGCCG	12.3	(4.0)
40	3e-24	Nfy(-)	ATTGGC	-100.0	(16.1)	ATTGGC	-67.8	(2.4)
41	2e-23	*m*12	GCAGCA	28.6	(15.4)	--	--	--
42	9e-23	Nfy(-)	TTGGCT	-70.9	(26.7)	TTGGCT	-69.7	(16.3)
43	3e-21	Inr	GGCAGT	-3.0	(0.2)	GGCAGA	-3.4	(0.6)
44	7e-21	*m*13	GGTGGC	45.1	(29.6)	--	--	--
45	2e-20	*m*14	GGACCC	102.1	(46.4)	GACCCC	72.2	(0.5)
46	9e-20	Inr	CAGTCY	-1.0	(0.2)	CAGTCTG	-1.3	(0.7)
47	9e-18	Inr	CACTTC	-1.0	(0.2)	--	--	--
48	1e-17	MEF2	AAAATA	202.3	(78.0)	--	--	--
49	1e-16	*m*15	GAAGGT	54.4	(38.3)	--	--	--

Table T3

Section S5: Supplementary MRF partner motif tables

Tables T4-T8

Partner motifs pairing with five predicted fixed motifs during the comprehensive MRF analysis. These include the NRF1 and EVI1 binding elements, as well as the novel y1-y3 motifs. Consensus motifs derived from each partner motif cluster are shown in the second column; factors binding to the known regulatory elements in TRANSFAC (4) are shown in the third column (STAMP (5) E-value threshold: 1e-5). Reverse complement matches for NRF1 are shown in the right-most column; numbers correspond to the rank-order on the left. E.g., the partner motif 2 is the reverse complement match of motif 20.

y1 partner motifs (TTTGTA)		
	Consensus	TF
1	TCCCAAAGTGCTG	TOPORS
2	AACTCCT	--
3	TTACAGG	--
4	ACCGC	--
5	TGAGCCA	JUN
6	ATGTTG	--
7	ATTAC	--
8	CGATCC	--
9	CTGAG	CDC5L
10	GCAAT	--
11	TGGGAT	PITX2
12	CTGGTCT	--
13	GTTGC	RFX
14	TCAAG	NKX2
15	TCCGC	--
16	TCTTG	EVI1
17	CTCGA	XBP1
18	GGCTG	--
19	TGTCA	TGIF

Table T4

NRF1 partner motifs (GCATGC)			
	Consensus	TF	RC
1	AGACG	--	
2	TAGAGA	--	[20]
3	GATTAC	--	
4	ATTCT	--	
5	CAAGC	--	
6	CGAACT	RXR	
7	CTGCC	--	
8	CTGGT	--	
9	GATCT	--	
10	GCAAC	RFX	
11	GGTCT	--	
12	GGTTTC	IRF1	[16]
13	GTAGA	--	[20]
14	GTATT	--	[17]
15	TTAGT	--	[19]
16	AAACC	--	[12]
17	AAATACAA	--	[14]
18	ACAAA	SOX10	
19	TACTAA	MEF2	[15]
20	TCTCTAC	--	[2]
21	TAAAA	TBP	
22	TCGAG	XBP1	
23	TGCAC	--	

Table T5

y2 partner motifs (ATTGC)		
	Consensus	TF
1	AAATTA	POU6F1
2	TACTAAAA	MEF2
3	AGCGA	--
4	CGTCT	--
5	CTCTAC	--
6	CTGTCT	SMAD3
7	TCAAA	TCF4
8	TCTCA	--

Table T6

EVI1 partner motifs (TCTTG)		
	Consensus	TF
1	TAATTT	POU6F1
2	TAGCTG	TOPORS
3	GCTAAT	CHX10
4	GTAGC	--
5	TTGTATTT	--
6	TCAGC	NFE2
7	TTTTA	TBP

Table T7

y3 partner motifs (GAGCT)		
	Consensus	TF
1	AAATACA	--
2	TACTAA	MEF2
3	ATCGA	CUX1
4	CAAAA	--
5	CAGCT	UBP1
6	CTCTAC	--
7	GCGAG	--

Table T8

Supplementary References

1. Broyden, C. (1965) A class of methods for solving nonlinear simulationeous equations. *Mathematics of Computation*, **19**, 577-593.
2. Davison, A.C. (2003) *Statistical Models*. Cambridge University Press, New York.
3. Vardhanabhuti, S., Wang, J. and Hannenhalli, S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res*, **35**, 3203-3213.
4. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**, 374-378.
5. Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, **35**, W253-258.
6. Maglott, D.R., Katz, K.S., Sicotte, H. and Pruitt, K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res*, **28**, 126-128.
7. Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, **29**, 137-140.
8. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.
9. Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol*, **7**, R78.
10. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, **38**, 626-635.