Supplementary Online Material

EVOLUTION OF THE MAMMALIAN TRANSCRIPTION FACTOR BINDING REPERTOIRE VIA TRANPOSABLE ELEMENTS

Guillaume Bourque*, Bernard Leong, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, Kandhadayar G. Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck Hui Ng and Edison T. Liu

*Correspondence: bourque@gis.a-star.edu.sg

Supplementary Text

Binding motif enrichment extends to homologous neighborhoods

Motif conservation across multiple species is a powerful way to identify *bona fide* transcription factor binding sites that play important regulatory roles (Wang et al. 2006). But of course, such an assessment will depend on our ability to correctly characterize the sequence motifs that are associated with the binding of the transcription factors. To begin addressing this, we measured the enrichment of predicted binding motifs in the ChIP-PET, ChIP-Chip and ChIP-Seq binding regions. The results for centered windows of size 200, 500, 1000 and 2000 bps are shown in Supplementary Figure 4. Overall, we find that in the smallest windows, the motif enrichment varies from 4.9 fold (ESR1-CC) to 50.5 fold (CTCF). An interesting finding is that although the incremental enrichment remains marginally significant in 500 bps windows it vanishes in larger windows (expect for TP53 which has an exceptionally low background). This attests to the accuracy and resolution of the 3 ChIP techniques. The fact that the motif enrichment is inversely correlated with the distance to the center of the binding regions can also be deduced from the distribution of the motifs within the regions (data not shown).

Next, relying on the strength of the motif enrichment, we used the same centered windows to look for motif conservation in other mammalian genomes. For the binding regions identified in human (ESR1, TP53, MYC, RELA), homologous regions in chimpanzee, macaque, mouse and dog were extracted using the tool liftOver and searched for cross-specie *conserved motifs*. For the binding regions identified in mouse (POU5F1-SOX2 and CTCF), a similar approach was used using homologous regions in rat, human and dog. The results displayed in Supplementary Figure 4 show that the fold enrichment for conserved motifs in 200 bps windows ranges from 14 fold for ESR1-CC to 190 fold for CTCF. The additional requirement of looking for motifs in other genomes implies a more stringent background and leads to overall stronger enrichments. What is more interesting is that incremental fold enrichment of conserved motif is no longer restricted to windows of small sizes (Supplementary Figure 4). For instance, using 1 Kbps windows instead of a 500 bps windows allows the recovery of 121 instead of 101 conserved RELA motifs. These 20 additional conserved motifs represent a 3.8 fold enrichment over the expected 5.26 ± 1.97 new conserved motifs and this difference is highly significant (p-value = 1.5E-12).

Overlap between the two conservation metrics

We were interested in looking at the agreement between the two conservation measures and the overall proportion of conserved binding sites for the different transcription factors. Supplementary Figure 5 displays the proportion of sites that contain a conserved motif only, a conserved element only, both a conserved motif and conserved element or neither. The most flagrant conclusion from this analysis is that the majority of binding regions are not conserved under either metric.

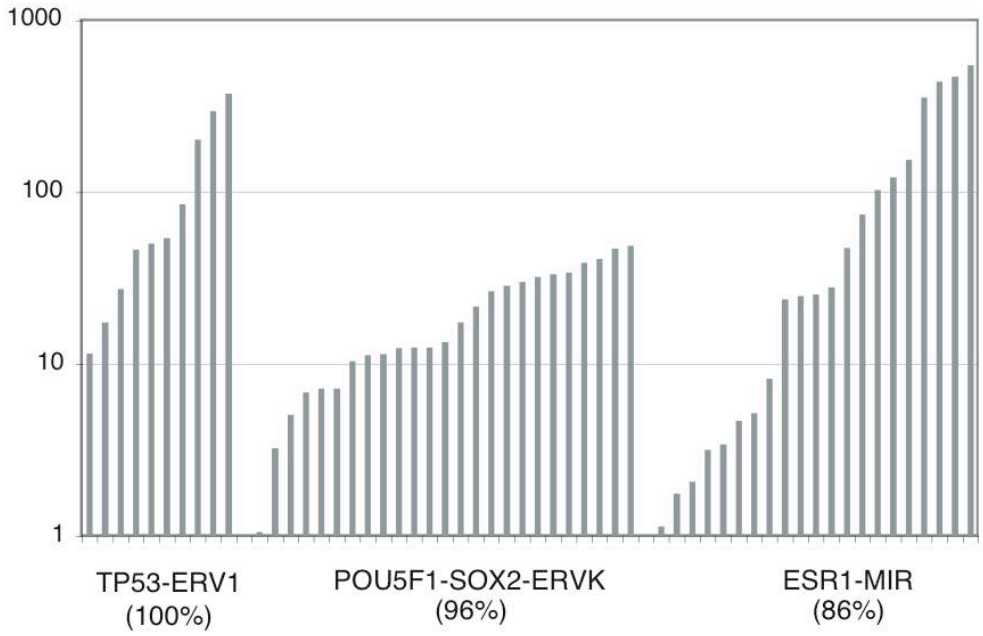
Association with Satellite and centr repeats

We found that 89 (13.5%), 39 (6.3%), 16 (4.8%) and 11 (0.9%) of the binding sites of MYC, RELA, TP53 and ESR1 respectively were associated with *Satellite* and *centr* repeats as compared to the expected 0.5%. The vast majority (127/155 = 82%) of these binding regions were in pericentromeric regions (i.e. within 5Mb of a centromere) and many were common across the different libraries. Because the current genome assembly is incomplete in pericentromeric regions which are known in particular to be depleted of satellite-rich sequences (She et al. 2004), we believe that the random genomic fragments coming from these regions are overrepresented and lead to misguided binding sites. This reduced binding potential is also corroborated by the lack of sequence binding motifs in those regions. For instance, although 34.5% of the non-satellite repeat MYC binding sites have an Ebox motif, this drops to 4.5% for MYC binding sites associated with satellite repeats. Similarly, the proportion of motif bearing RELA binding sites goes from 48.3% to 0% in the 39 sites that are associated with this class of repeats. For these reasons, we have removed these particular sites from the downstream analyses.

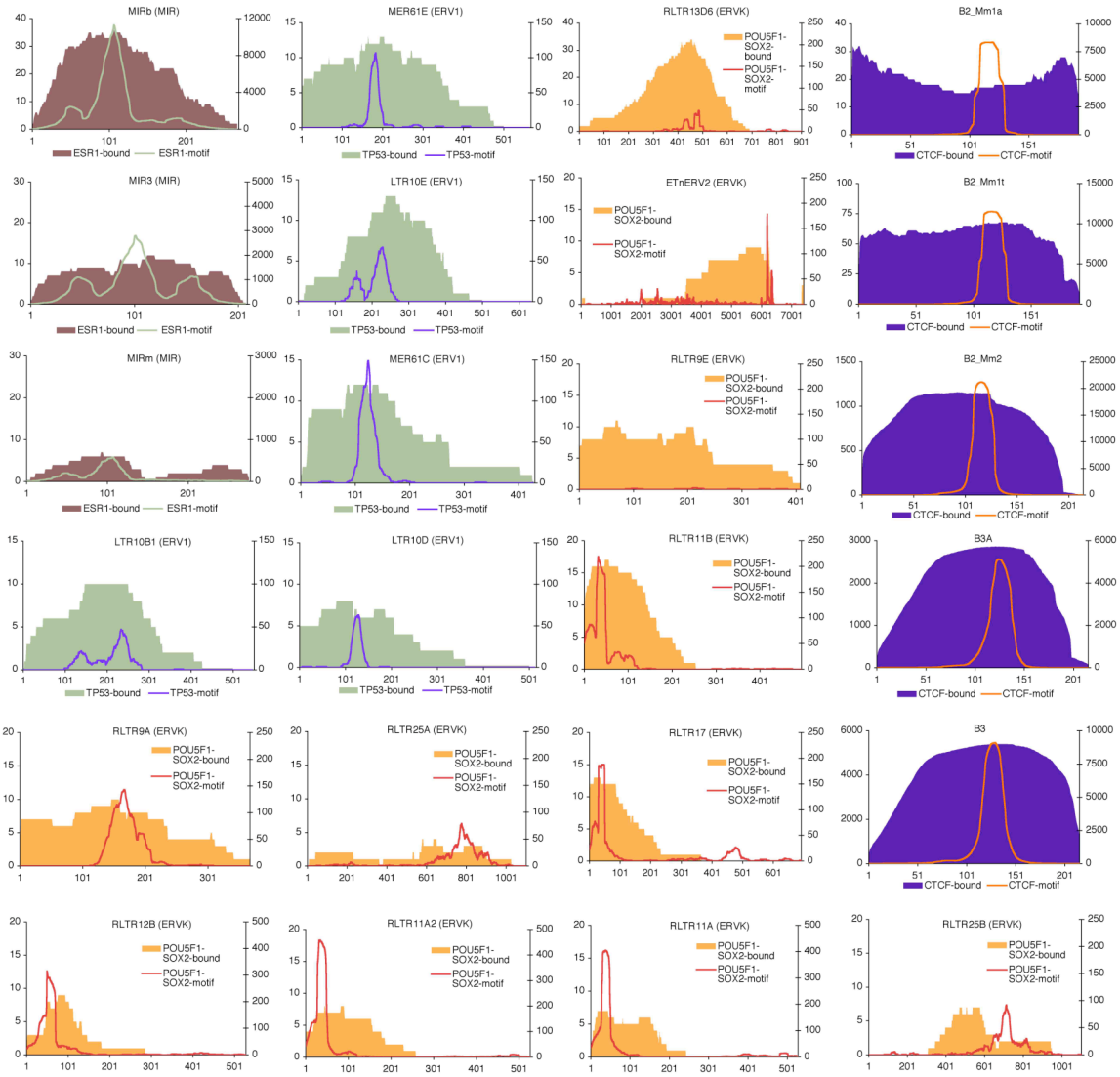
ESR1 RABS in ChIP-PET and ChIP-Chip data sets

When assessing the differences between the output of estrogen receptor (ESR1) binding site maps derived from ChIP-PET (a sequence based assessment) and from ChIP-Chip (a hybridization based assessment) platforms, we observed that there was a significant difference in the identification of binding sites that reside in repeats. The arrays used for ChIP-Chip experiments routinely mask repetitive sequences in the probe regions, whereas the sequence-based assessment in the ChIP-PET strategy is free of this constraint. Thus, although 18% of the ChIP-PET determined ESR1 binding sites contained traces of the MIR repeat, only 11% of the ChIP-Chip binding sites had the same repeat (data not shown).

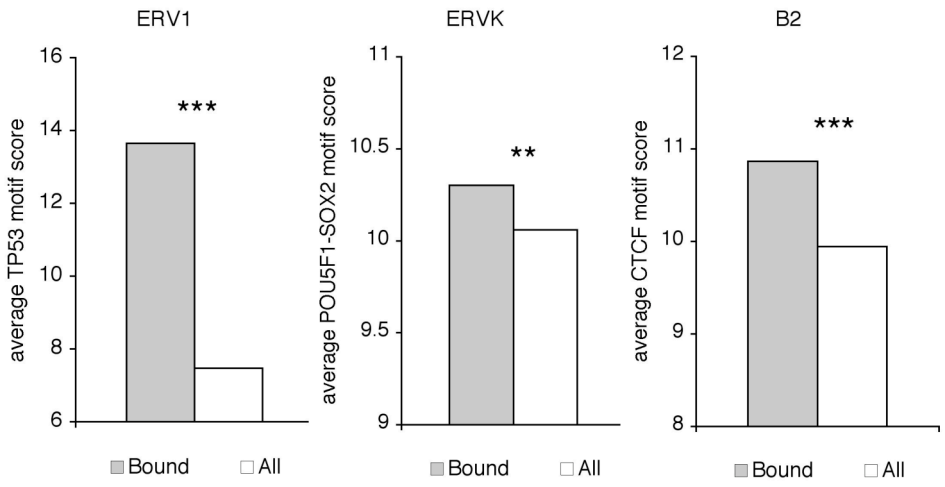
Supplementary Figures



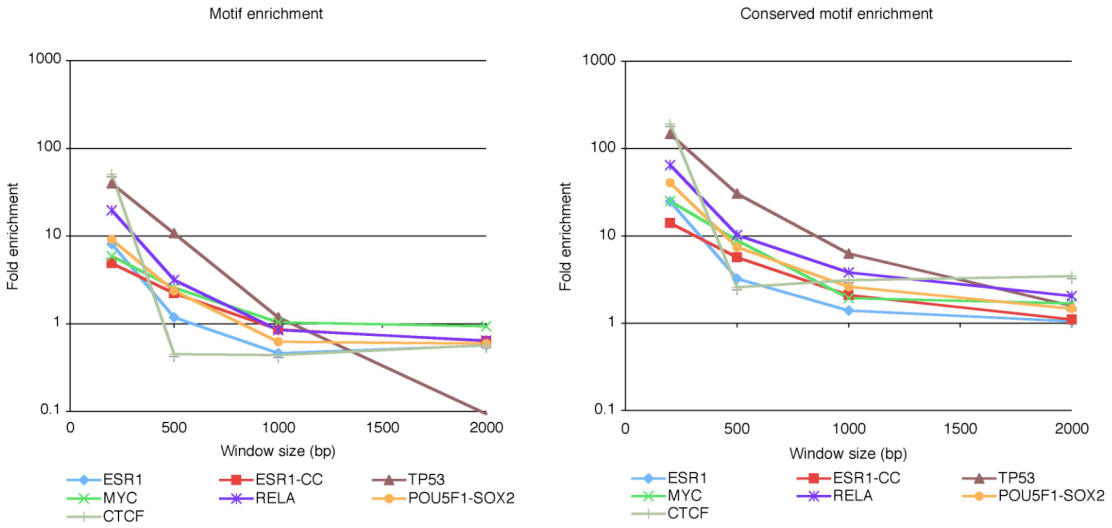
Supplementary Figure 1. ChIP-qPCR fold enrichments of repeat-associated binding sites for TP53, POU5F1-SOX2 and ESR1. Validation rates are computed using a minimum fold-enrichment of 3.



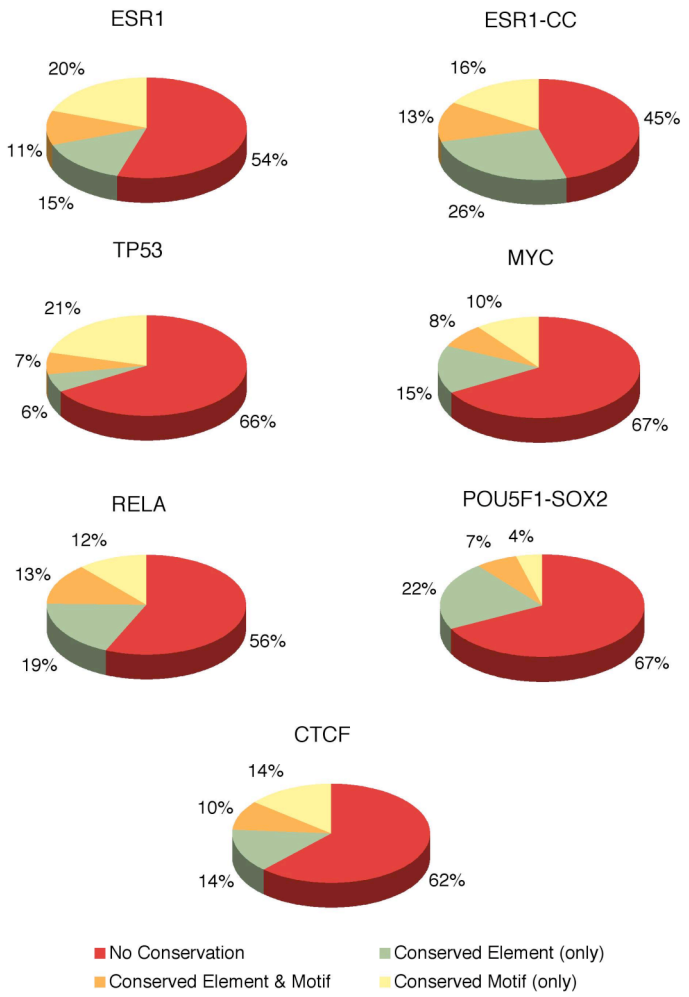
Supplementary Figure 2. Expanded version of Fig. 3A from main text including all over-represented families from Table 1 in the main text.



Supplementary Figure 3. Bound transposable elements have better motif instances than the unbound ones. P-values are based on 1000 random samples: *** implies p-value < 0.001 while ** implies p-value < 0.01.



Supplementary Figure 4. Incremental motif enrichment obtained in windows of increasing size (200, 500, 1000 and 2000 bps) centered on the middle of the binding regions identified for ESR1, TP53, MYC, RELA, POU5F1-SOX2 and CTCF. The incremental enrichment is the number of additional motifs detected in a particular window size divided by the expected number of such newly detected motifs. In 200 bps windows all motifs are considered to be new.



Supplementary Figure 5. Percentage of binding regions overlapping a conserved element (grey), with a conserved motif (yellow), both a conserved element and motif (orange) or with no sequence conservation (red).

Supplementary Tables

	Binding sites	Conserved Element (%)	Conserved Motif (%)	Non-conserved (%)
ESR1	1234	25.9	30.6	54.6
ESR1-CC	3665	38.1	28.8	45.7
TP53	336	12.8	27.7	66.4
MYC	660	23.2	18.0	66.7
RELA	617	31.9	24.6	56.6
POU5F1-SOX2	1507	28.4	10.8	67.5
CTCF	39609	24.0	24.0	61.9

Supplementary Table 1. A majority of binding regions do not show signs of sequence conservation based on either overlap to PhastCons Conserved Elements or presence of conserved binding motifs.

	Binding motif (%)	RABS and Binding motif (%)
ESR1	61.1	60.7
TP53	84.2	97.2
POU5F1-SOX2	44.1	55.9
CTCF	68.8	78.3

Supplementary Table 2. Percentage of binding regions and repeat-associated binding regions (RABS) that have a binding motif.

	Repeat Name	Hamming distance of the best site	Promoters with comparable or better motif (%)
ESR1	MIRb	0	8.7
	MIR3	1	45.3
	MIRm	0	9.0
TP53	MER61E	0	3.9
	LTR10E	2	95.9
	MER61C	0	2.8
	LTR10D	1	33.8
	LTR10B1	0	3.7
POU5F1-SOX2	RLTR13D6	2	97.7
	ETnERV2	1	99.5
	RLTR9E	3	98.2
	RLTR11B	0	2.9
	RLTR17	0	6.6
	RLTR9A	2	80.2
	RLTR12B	0	3.4
	RLTR11A2	0	3.4
	RLTR11A	0	3.4
	RLTR25B	0	9.6
	RLTR25A	0	10.2
CTCF	B2_Mm1a	0	3.6
	B2_Mm1t	0	3.6
	B2_Mm2	0	3.7
	B3	0	3.7
	B3A	0	4.0

Supplementary Table 3. Consensus sequences of the bound repeats were found to be better progenitors of binding motifs. We measured the hamming distance of the best site in each repeat consensus (as a proxy of the minimum mutational events required to generate a binding motif) and calculated the fraction of promoters in the whole genome that contain similar or closer sequences to the binding motif.

	Repeat Subfamily	Age (Myrs) Jukes Cantor	Age (Myrs) Kimura	Age (Myrs) PAML (w/o GC)	Age (Myrs) PAML (with GC)
SINE	MIR3	172	166	130	135
	MIRb	170	168	126	136
	MIRm	163	161	127	134
ERV1	LTR10B1	96	92	78	81
	LTR10D	57	71	46	56
	LTR10E	65	75	57	64
	MER61C	78	68	59	65
	MER61E	84	85	60	68
ERVK	ETnERV2	47	38	29	32
	RLTR9A	52	50	39	43
	RLTR9B	37	36	27	32
	RLTR11A	52	47	43	44
	RLTR11A2	55	53	40	46
	RLTR11B	52	49	40	43
	RLTR12B	67	60	31	53
	RLTR13D6	41	39	31	34
	RLTR17	50	47	33	40
	RLTR25A	53	48	38	41
	RLTR25B	54	48	39	42
B2	B2_Mm1a	13	15	11	13
	B2_Mm1t	19	22	16	19
	B2_Mm2	25	27	22	24
	B3	60	61	44	51
	B3A	69	68	49	57

Supplementary Table 4: The age of a repeat subfamily computed using the RepeatMasker data using three methods: (i) Jukes Cantor using the divergence statistic in RepeatMasker, (ii) Kimura 2-distance using the transitions and transversions in RepeatMasker, and (iii) Divergence computed from PAML using sequence data with and without masking GC content.

	Repeat Name	Nb repeats	Observed motifs / repeat	Expected motifs / repeat	p-value
ESR1	MIRb*	280513	0.058	0.095	1
	MIR3*	72027	0.050	0.060	0.964
	MIRm*	32126	0.017	0.046	1
TP53	MER61E	320	0.400	0.428	0.961
	LTR10E	253	0.672	0.013	< 0.001
	MER61C	288	0.674	0.579	< 0.001
	LTR10D	190	0.395	0.063	< 0.001
	LTR10B1	238	0.563	0.418	< 0.001
POU5F1-SOX2	RLTR13D6	1239	0.091	0.018	< 0.001
	ETnERV2	4491	NA	NA	NA
	RLTR9E	1402	0.005	0.001	0.003
	RLTR11B	1944	0.208	0.138	< 0.001
	RLTR17	2642	0.156	0.095	< 0.001
	RLTR9A	1652	0.184	0.006	< 0.001
	RLTR12B	907	0.684	0.238	< 0.001
	RLTR11A2	3101	0.236	0.121	< 0.001
	RLTR11A	2897	0.228	0.091	< 0.001
	RLTR25B	4322	0.099	0.046	< 0.001
	RLTR25A	3179	0.142	0.069	< 0.001
CTCF	B2_Mm1a*	17753	0.487	0.477	0.164
	B2_Mm1t*	22203	0.535	0.511	0.015
	B2_Mm2*	85463	0.278	0.299	0.989
	B3*	140073	0.058	0.096	1
	B3A*	87707	0.057	0.065	0.928

Supplementary Table 5. Repeat instances are significantly enriched for binding motifs. The computation of expected motif was carried out through 1000 Monte Carlo simulation of random mutations, taking into account the length and amount of substitution in each repeat instance. *For these large families, the analysis is based on a sampling of 2000 repeat instances.

	Repeat name	Total motifs (nb)	Bound motifs (nb)	Bound motifs (%)
ESR1	MIR3	7185	15	0.21
	MIRb	24202	46	0.19
	MIRm	1433	6	0.42
	MIR	48869	91	0.19
	Random (1M)	30104	53	0.18
TP53	MER61E	142	22	15.49
	LTR10E	171	41	23.98
	MER61C	199	17	8.54
	LTR10D	75	11	14.67
	LTR10B1	137	21	15.33
	ERV1	9001	143	1.59
	Random (1M)	6053	10	0.17
POU5F1-SOX2	RLTR13D6	116	4	3.45
	ETnERV2	1189	12	1.01
	RLTR9E	9	0	0.00
	RLTR11B	428	14	3.27
	RLTR17	413	7	1.69
	RLTR9A	310	6	1.94
	RLTR12B	621	8	1.29
	RLTR11A2	735	6	0.82
	RLTR11A	661	6	0.91
	RLTR25B	458	6	1.31
	RLTR25A	454	11	2.42
	ERVK	16145	104	0.64
	Random (1M)	12609	25	0.20
CTCF	B2_Mm1a	8684	14	0.16
	B2_Mm1t	12057	59	0.49
	B2_Mm2	22971	993	4.32
	B3A	6438	2200	34.17
	B3	11309	4050	35.81
	B2	61459	7316	11.90
	Random (1M)	4491	353	7.86

Supplementary Table 6. Repeat instances are enriched for bound motifs. Expected levels where measured in a sample of 1 million random positions from the corresponding genome.

POU5F1-SOX2 RABS	Gene name	Affymetrix ID	Relative Position	Distance
chr7.114695151	Sept1	1449898_at	5'	5999
chr11.106395678	Pecam1	1421287_a_at	5'	9451
chr3.89063382	Ubqln4	1448691_at	3'	2392
chr5.75600278	A730073F16Rik	1419824_a_at	3'	6617
chr1.17034092	NA	1437867_at	3'	7450
chr17.27381418	Mapk13	1448871_at	3'	8505
chr17.45878480	Frs3	1424449_at	3'	9124
chr12.4786805	Ubxd4	1425020_at	inside	8914
chr2.160719756	Top1	1423474_at	inside	10558
chr10.6059645	Akap12	1419706_a_at	inside	23858
chr10.61111717	X99384	1448134_at	inside	32474
chrX.67386874	Pls3	1423725_at	inside	44506
chrX.21651446	Klhl13	1448269_a_at	inside	45939
chr3.136093140	Manba	1450626_at	inside	60936
chr1.164394053	Nme7	1418217_at	inside	69074
chr4.144685145	Rex2	1426137_at	inside	70612
chr10.108082996	Pawr	1426910_at	inside	71287
chr1.34386946	Dst	1423626_at	inside	92177

Supplementary Table 7. Repeat associated POU5F1-SOX2 binding sites within 10Kb of an POU5F1 or SOX2 regulated genes from (Ivanova et al. 2006).

ESR1 RABS	Gene Name	Affymetrix ID	Relative Position	Distance
chr19.52531896	GPR77	221149_at	5'	115
chr10.51217192	MSMB	207430_s_at	5'	2063
chr9.129597065	PTGES	210367_s_at	5'	2829
chr9.92913871	SUSD3	227182_at	5'	3619
chr20.48776387	PARD6B	211907_s_at	5'	4203
chr5.139002858	CXXC5	224516_s_at	5'	4836
chr15.61465657	CA12	203963_at	5'	6048
chr20.45555301	NCOA3	209060_x_at	5'	7800
chr3.50617320	CISH	223961_s_at	3'	1517
chr12.96464503	NCRMS	229782_at	3'	3852
chr3.151937578	SIAH2	209339_at	3'	3998
chr16.8787729	ABAT	206527_at	3'	5041
chr9.4855810	RCL1	218544_s_at	3'	5383
chr17.55279697	VMP1	224917_at	3'	5518
chr1.203456894	TOSO	221601_s_at	3'	8884
chr4.89597068	HERC6	244760_at	inside	2359
chr20.57995993	CDH26	232306_at	inside	5416
chr20.52106704	BCAS1	204378_at	inside	10265
chr3.14429456	SLC6A6	205921_s_at	inside	10467
chr17.70268412	SLC9A3R1	201349_at	inside	12213
chr20.34646533	TGIF2	218724_s_at	inside	13780
chr16.8678085	ABAT	206527_at	inside	15240
chr4.89612975	HERC6	244760_at	inside	18400
chr15.69374273	FLJ13710	222835_at	inside	19746
chr8.11625631	GATA4	205517_at	inside	22806
chr11.35366420	SLC1A2	208389_s_at	inside	28973
chr1.21710558	RAP1GA1	203911_at	inside	30297
chr11.30499246	C11orf8	205413_at	inside	58681
chr20.19212066	SLC24A3	219090_at	inside	70780
chr14.88004385	PTPN21	1320_at	inside	81200
chr5.142679813	NR3C1	201865_x_at	inside	82602
chr20.52030817	BCAS1	204378_at	inside	87366
chr11.30416645	C11orf8	205413_at	inside	141647
chr17.56663045	BCAS3	220488_s_at	inside	241182
chr17.56765385	BCAS3	220488_s_at	inside	343904
chr17.56797790	BCAS3	220488_s_at	inside	377662
chr16.77628669	WWOX	219077_s_at	inside	938164

Supplementary Table 8. Repeat associated ESR1 binding sites within 10Kb of an ESR1 regulated genes from (Lin et al. 2007).

Supplementary Table 9. ESR1 binding sites (see file Table_S9_ESR1_binding.tsv).

Supplementary Table 10. ESR1-CC binding sites (see file Table_S10_ESR1CC_binding.tsv).

Supplementary Table 11. TP53 binding sites (see file Table_S11_TP53_binding.tsv).

Supplementary Table 12. MYC binding sites (see file Table_S12_MYC_binding.tsv).

Supplementary Table 13. RELA binding sites (see file Table_S13_RELA_binding.tsv).

Supplementary Table 14. POU5F1-SOX2 binding sites (see file Table_S14_O4S2_binding.tsv).

Supplementary Table 15. CTCF binding sites (see file Table_S15_CTCF_binding_wBarski.tsv).

Supplementary References

Ivanova, N., R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I.R. Lemischka. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442:** 533-538.

Lin, C.Y., V.B. Vega, J.S. Thomsen, T. Zhang, S.L. Kong, M. Xie, K.P. Chiu, L. Lipovich, D.H. Barnett, F. Stossi et al. 2007. Whole-Genome Cartography of Estrogen Receptor alpha Binding Sites. *PLoS Genet* **3:** e87.

She, X., J.E. Horvath, Z. Jiang, G. Liu, T.S. Furey, L. Christ, R. Clark, T. Graves, C.L. Gulden, C. Alkan et al. 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430:** 857-864.

Wang, H., Y. Zhang, Y. Cheng, Y. Zhou, D.C. King, J. Taylor, F. Chiaromonte, J. Kasturi, H. Petrykowska, B. Gibb et al. 2006. Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res* **16:** 1480-1492.