SUPPLEMENTAL METHODS

Development of a statistically-based peak-calling method. We began by considering, based on our experimental parameters, what type of signal would be produced by a genuine binding site in this representation of high-density oligonucleotide data. The DNA placed on the arrays averaged ~300 nts in length. Given the tiling parameters of the array and the fact that true binding sites are expected to be <50 nts in length, a single binding site would ideally be represented by a "hill"-like waveform in the data of length 2 x 300=600 nts. However, in actuality, we expected that this ideal, simple hill-like waveform will rarely be observed for the following reasons: (1) some "spots" may hybridize better than others, producing jagged waveforms; (2) several binding sites could be close to each other, producing elongated peaks without single hill-like shape; and (3) repeat-masking on the arrays can lead to "missing points", producing array waveforms with interruptions and even premature termination. Indeed, inspection of a "gallery" of peaks (all of which have been confirmed using standard PCR; data not shown), demonstrates that peaks corresponding to confirmed binding sites vary greatly in waveform, amplitude, and size (**Supplemental Figure S1B**). Furthermore, we find that there is not always agreement between biological replicates, even after scaling of arrays. Supplemental Figure S1C1 shows an example in which arrays are scaled to emphasize the similarity of the large peak in box b; clearly, there is a smaller peak (box a) that is present in array A but not in the other arrays. Also, in other cases, some peak-like waveforms were present in two of three arrays (**Supplemental Figure S1C2**). Hence, at least some array differences are not simply scaling effects but reflect real differences in the results.

For our peak detection method, we use a percentile for each array (95th and 98th percentile) of log2 oligomer ratios. For a given threshold, we recode the array with points above the threshold as 1, points below as 0, and missing regions (due primarily to repeat-masking) as X **(Supplemental Figure S2**). Because we have observed that some confirmed binding sites display peak-like waveforms that are "interrupted" by missing points, we ignore the X points for our analysis. This results in a long string of ~380,000 1s and 0s to represent the full set of oligomers for the ENCODE regions on the array. Genuine binding sites should be represented as a series of points of elevated amplitude; this is equivalent to a run of 1s. However, just by chance there will occasionally be two or more consecutive points of high amplitude that will be coded as 1s. Hence, we wish to set our width requirement for detecting a genuine binding site to be greater than what would appear randomly. Given the sequence length (e.g. ~380,000 for our ENCODE arrays) and the probability of getting a 1 at any position, we can calculate the longest run of 1s expected purely by chance using the well-known Erdos-Renyi Law (Erdos and Renyi 1970). Importantly, this has been extended to exact results enabling calculation of the actual probabilities (p-values) associated with each run length (Waterman et al. 1987). We use the following 6 relations (Waterman et al. 1987) to calculate p-values for a run of 1s, where w is the length of the run; L is the length of the sequence (~380,000 for these ENCODE arrays); p is the probability of having any given point be a one (for 98th percentile threshold, this value is (1-0.98)=0.02); SD denotes the standard deviation; z is the z-score; and P is the final p-value reflecting the probability of getting this run length of 1s:

(1) mean $R_n = \log_{(1/p)} L + (0.577)/\Theta - 0.5$

(2) variance $R_n = \pi^2/(6\Theta^2) + 1/12$

(3) $\Theta = \ln 1/p$

(4) $SD = \sqrt{\text{variance}}$

(5) $z = (w - \text{mean } R_n)/SD$

(6) $P(Z>z) = 1-\exp(-\exp(-1.2825z - 0.577))$

Following Waterman et al. (1987), the first two of these indicate the mean and variance for the maximum run length; the third is a convenient definition; the fourth and fifth are listed for clarity and reflect well-known formulae for the standard deviation and z-score; the last indicates the equation for conversion of z-score to p-value (based on extreme-value distribution).

For each threshold (95th percentile and 98th percentile), we use $p<0.0001$ for a very stringent p-value and $p<0.05$ for a less stringent p-value cutoff. For our ENCODE arrays of ~380,000 points, $p<0.0001$ requires 6 consecutive points above the 98th percentile or 8 consecutive points above the 95th percentile, while $p<0.05$ requires 4 consecutive points above the 98th percentile or 5 consecutive points above the 95th percentile. The combination of a high threshold and small p-value is obviously more stringent than a low threshold and larger p-value. Hence, our four conditions, in decreasing stringency are: 98th percentile threshold and $p<0.0001$; 95th percentile threshold and $p<0.0001$; 98th percentile threshold and $p<0.05$; 95th percentile and $p<0.05$. For clarity, we refer to these as L1-L4, with L1 being the most stringent. It is important to note that as we lower stringency, we keep adding peaks to the set. So (with a few very rare exceptions – see below section), every L1 peak is present at L2; every L2 peak is present at L3, etc. Hence, we will distinguish the *set of L2 peaks* (meaning every peak present in the L2 set) from *peaks that first appear at L2* (which is the "set of L2 peaks" minus "the set of L1 peaks"). It is important to note that this formulation allows us to analyze the data from each array in a consistent and unbiased manner. For example, to compare arrays at L1, for each array separately, we calculate the 98th percentile threshold and use a minimum run length of 6 to yield peaks with $p<0.0001$.

Rationale for Peak-First Strategy and Combination Procedure

We chose the "peak-first" strategy (Figure 3) because (1) when one array has much stronger signals than the others, the "strong" array could over- influence the peak predictions in the combined data set; (2) array combination procedures implicitly assume that each point is completely independent of all others, hence these procedures might upset the "runs" of points that our procedure relies on for binding site detection; (3) this "peak-first" combination approach tended to avoid cases in which a barely subthreshold point would interrupt a peak. Therefore, we call the peaks on each array and then designate a "binding site" as a region that is called as a peak on two of the three arrays. To combine the peak predictions from the separate arrays into a single binding site, we take the union of peak predictions that are overlapping or very close to overlapping (peaks separated by < 100 nt were considered overlapping) to compensate for the occasional 50 mer probe that was below threshold (**Figure 3C**).

Relationship between sets of peaks at L1-L4.

There are three significant considerations here. First, at L1, the threshold is the 98th percentile and a run of 4 consecutive points is required. At L2, the threshold is the 95th percentile and a run of 8 consecutive points is required. Hence, a run of 4-7 points at the 98th percentile would be detected as an L1 peak but would not be present in the L2 set. In practice, this situation occurred rarely; >96% of L1 peaks are in the L2 set. Second, it is important to note that, by definition, every L1 peak is present in the L3 set and every L2 peak is in the L4 set. This follows from the fact that L3 uses the same threshold as L1, except a smaller minimum run length. Similarly, L4 uses the same threshold as L2 except that a shorter run length is used. Third, we consistently observe that the 95th percentile/$p<0.0001$ (we refer to as L2) and 98th percentile/$p<0.05$ (we refer to as L3) yield very similar results (see Supplemental Table S2) in the data sets in this study and other data sets for other factors (S. Squazzo, A. Rabinovich, personal communications). We have chosen the 95th percentile/$p<0.0001$ as L2 and 98th percentile/$p<0.05$ as L3 because we also consistently observe more predicted binding sites at 98th percentile/$p<0.05$ than at 95th percentile/$p<0.0001$.

Calculation of predicted number of points in promoter array corresponding to real E2F1 binding sites for the promoter array

There are ~360,000 oligomers on the promoter array, representing ~24,000 promoters, with each promoter being represented by 15 points with an average spacing of

100 nts. ENCODE data suggests that E2F1 regulates ~25% of transcripts (35% of genes; see RESULTS), yielding a conservative approximate expectation of 0.25*24,000=6,000 promoters bound. Now, each of these 6,000 is represented by 15 points. If 5/15 points are actually in an E2F1 binding site (a conservative expectation given inspection of the raw data), then there are 6,000 x 5= 30,000 points actually in E2F1 binding sites. Hence, 30,000/360,000=(1/12) meaning that 1 of every 12 points in the array may be bound by E2F1. Changes in the assumptions here will only shift these numbers by a small amount.

Quantification of PCR results

For the great majority of PCR confirmation experiments, the relationship between ChIP and Total samples could be easily ascertained via simple visual inspection. To test the reliability of these calls made by visual inspection, we quantitated gel image intensities (Quantity One Software, Bio-Rad Inc., Hercules, CA) and normalized by the ratio of E2F1/Total for a known negative control using each amplicon set (A, B, C sets as discussed in RESULTS) (Chr21 control; same primers as used in Fig 1 for ChIP verification) for a subset of experiments at L1 (n=13 peak predictions, representing 39 (=13 x 3 biological replicates) total array predictions). Using this quantitation approach, all 13 peaks were confirmed (>=2 of 3 replicates show enhancement) and 38/39 (>97%) of individual array visual predictions were confirmed. Hence, visual inspection of the PCR data appeared highly reliable.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure S1. Variability of Raw Data and Peak Forms. (A) The data for ENCODE region ENr232 from three different E2F1 ChIP experiments are shown to emphasize the significantly different properties of the binding patterns with respect to maximal amplitudes and apparent noise that are observed in different samples. The x axis indicates the position along the chromosome whereas the y axis indicates the ratio of the signal from the E2F1 ChIP sample divided by the total sample for each oligomer (in log2). (B) A gallery of peaks corresponding to genuine E2F1 binding sites. Each ratio axis ranges from –1 to 3 (in log2); each bottom axis is 5 kb. Although there is great variability in waveform and amplitude, each of these peaks was confirmed as an E2F1 binding site by PCR using the amplicons that were applied to the array. (C) Some peak-like waveforms are not present in every array. (1) Changing the scaling of the arrays can show similar peaks in some cases, e.g. the peak shown in Box b. However, in Box a, the clear peak in array A is not present in the other arrays. (2) An example of a peak waveform present in array A and array C but not array B. Note that array B has negative-going signal whereas arrays A and C show a clear peak-like waveform.

Supplemental Figure S2. Peak Identification. A segment of data from ENr333 is shown in detail. The two dotted lines show the values for the 95th percentile threshold (lower dotted line) and the 98th percentile threshold (upper dotted line), which were computed for the whole array. The computed string after thresholding the data in the box for the 98th percentile level is shown, where 1 indicates a point greater than or equal to the 98th percentile threshold, 0 indicates a point below the threshold, and x indicates a section of missing oligomers (which generally are due to repeat masking). The run of five '1's in the middle of the string would be detected as a peak at the “L3” level (see RESULTS); this peak was confirmed as a binding site via PCR.

Supplementary Figure S3. Analysis of E2F1 target genes identified on the human promoter array. (A) Predicted E2F1 Target Promoters. PCR confirmations of 10 randomly chosen promoters (from the set of promoters with median values > 0.707; see RESULTS) that were identified as E2F1 target promoters using the human promoter array. (B) Known

E2F1 Target Promoters. Examples of the binding patterns for 11 different known E2F target genes as obtained from the ChIP-chip assay using the human promoter array. The binding pattern of a promoter (*Neu4*) that was not identified as a target promoter is shown for comparison. The log2 value of the maximum amplitude bar, and the median value of all bars for that promoter is shown in each box. Note that although the promoter array is based on 15 oligos/1500 nt promoter region (represented as 15 bars/region), some displayed regions (e.g. *Dhfr*) have 30 bars because there are two closely spaced promoters present (i.e. bidirectional promoter regions). All gene names are HUGO names except for two genes not included in the HUGO database: ENSG00000196112, which is the ENSEMBL gene name, and AF116678, which is the EMBL mRNA name (no gene name available).

Supplementary Figure S4. Great Majority Of E2F1 Binding Sites Are In CpG Islands And Lack The Consensus Binding Site Motif. (A) A portion of region ENr333 (chr20) is shown with the vertical bars at each track indicating the location of E2F1 binding sites predictions at L1, CpG islands, TTTSSCGC, and TTTSSCGC in which one T is allowed to be any nucleotide (TTTm1SSCGC). Both strands were examined for motif locations. Shaded vertical boxes highlight each experimentally determined E2F1 binding site. (B) Overlap of E2F1 binding sites with TTTSSCGC motifs. (C) Overlap of E2F1 binding sites with CpG islands. (D) Weblogo of the 27 TTTSSCGC motifs found in the experimentally determined E2F1 binding sites.

Supplementary Figure S5. Localization Of MYC, E2F1, And RNA Polymerase II (POLR2A) Binding Sites In A Portion Of ENCODE Region ENm005 (Chr21). Vertical shaded regions indicate binding sites that are near 5' ends of genes in this region. Note that *Ifnar1* has binding sites for MYC, E2F1, and POLR2A near the 5' end, that all E2F1 binding sites are near 5' ends, and that a subset of the E2F1 sites show bound POLR2A. Only one MYC binding site is near the 5' end of a gene; the arrows indicate the MYC binding sites that are far from 5' ends. All binding sites are L1 stringency sites (see RESULTS). Reworked image from UCSC genome browser (hg16) shows all genes but only one splice variant/gene for presentation clarity. Gene set is Known Genes II set (June 2005).

Supplemental Figure S6. MA plots of HeLa E2F1 array data.
Plots are for A, B, C arrays (as referred to in text) from top to bottom. Because each plot encompasses ~380,000 points, and the finite minimum size of pixels in this representation, the "smear" of points does not represent adequately the actual density of points in these plots, and comparison of different plot regions that appear completely black are inherently misleading. To illustrate this, we chose three equal areas in the solid black region of the top plot - these are marked with white rectangles. Moving left to right, we find that there are 12281, 9127, and 1315 points in each. Hence, although solid black areas appear to be identical on the plot, they can vary by at least 900% in actual point counts, and hence other quantitative analyses must be used in lieu of direct examination of these plots. The original data sets should be consulted for more detailed analysis. The lines on the plots indicate the 95th and 98th percentile values for each.

Figure S1

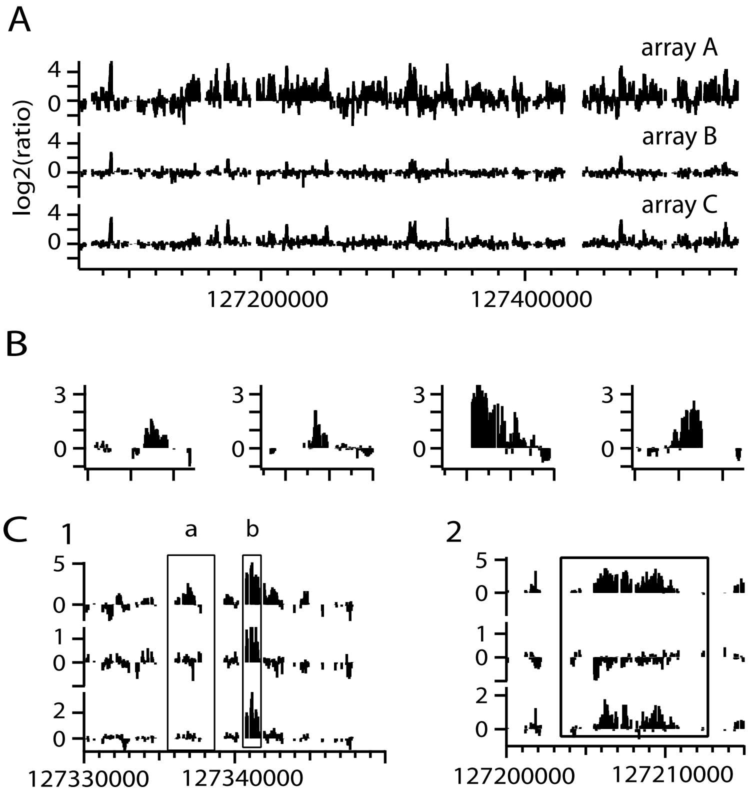

A
array A
array B
array C
log2(ratio)
4
0
4
0
4
0
127200000
127400000
B
3
0
3
0
3
0
3
0
C
1
a
b
5
0
1
0
2
0
127330000
127340000
2
5
0
1
0
2
0
127200000
127210000

Figure S2

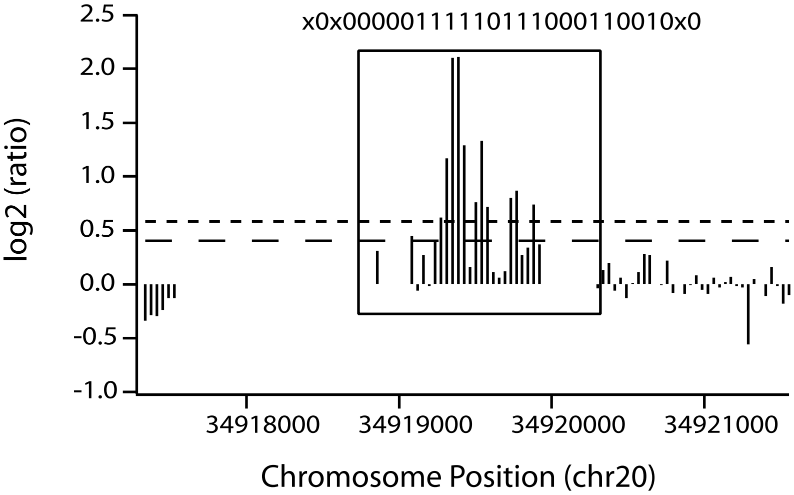

x0x000001111101111000110010x0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
log2 (ratio)
34918000
34919000
34920000
34921000
Chromosome Position (chr20)

Figure S3

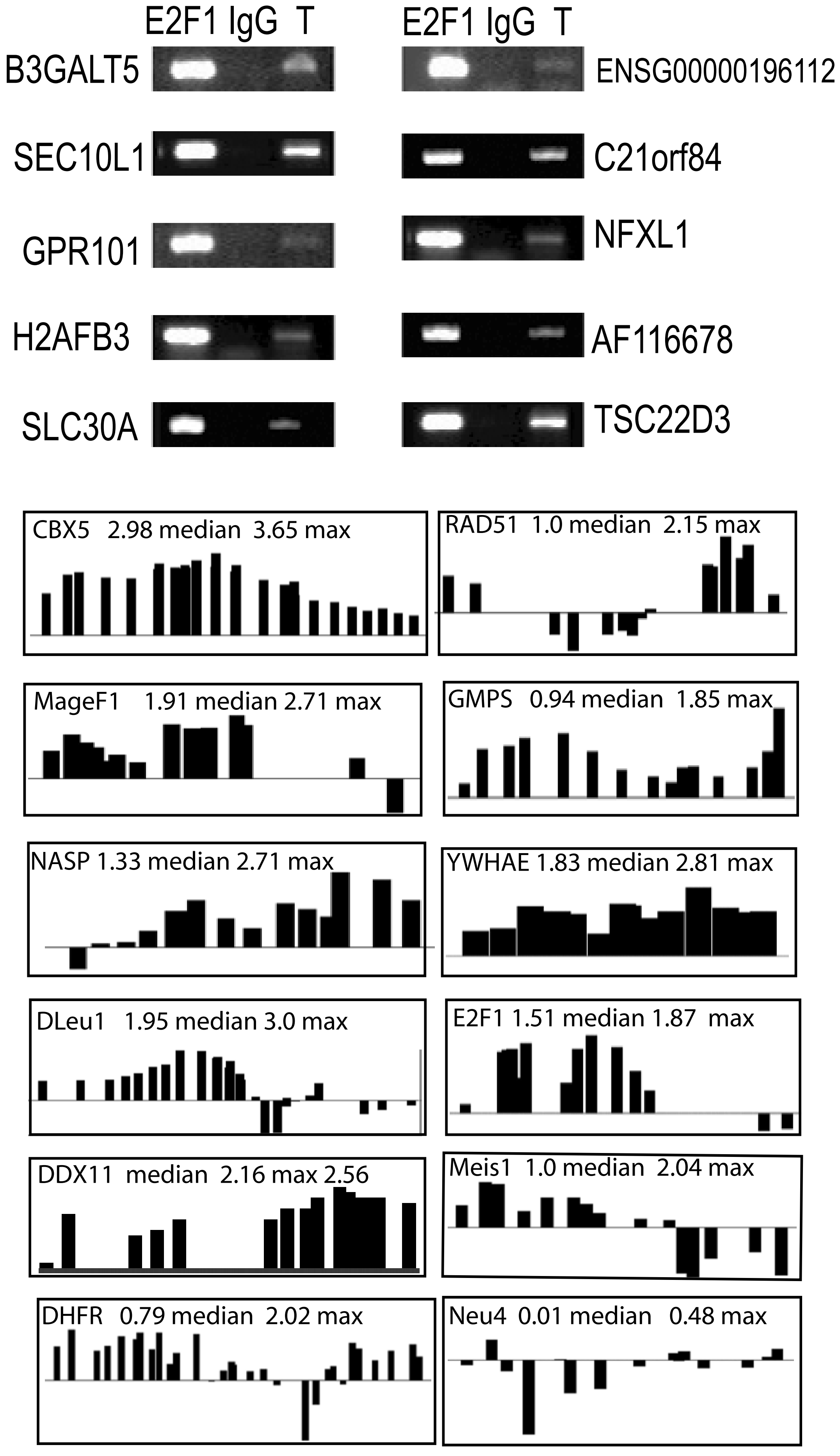

E2F1 IgG T
E2F1 IgG T
B3GALT5
ENSG00000196112
SEC10L1
C21orf84
GPR101
NFXL1
H2AFB3
AF116678
SLC30A
TSC22D3
CBX5 2.98 median 3.65 max
RAD51 1.0 median 2.15 max
MageF1 1.91 median 2.71 max
GMPS 0.94 median 1.85 max
NASP 1.33 median 2.71 max
YWHAE 1.83 median 2.81 max
DLeu1 1.95 median 3.0 max
E2F1 1.51 median 1.87 max
DDX11 median 2.16 max 2.56
Meis1 1.0 median 2.04 max
DHFR 0.79 median 2.02 max
Neu4 0.01 median 0.48 max

Figure S4

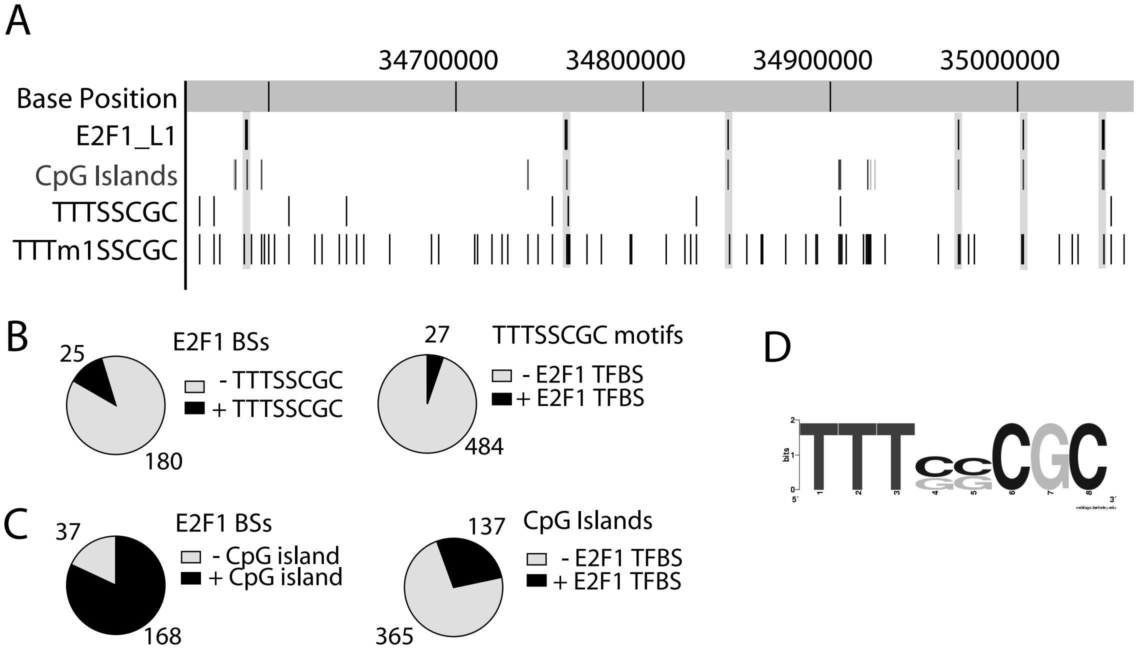
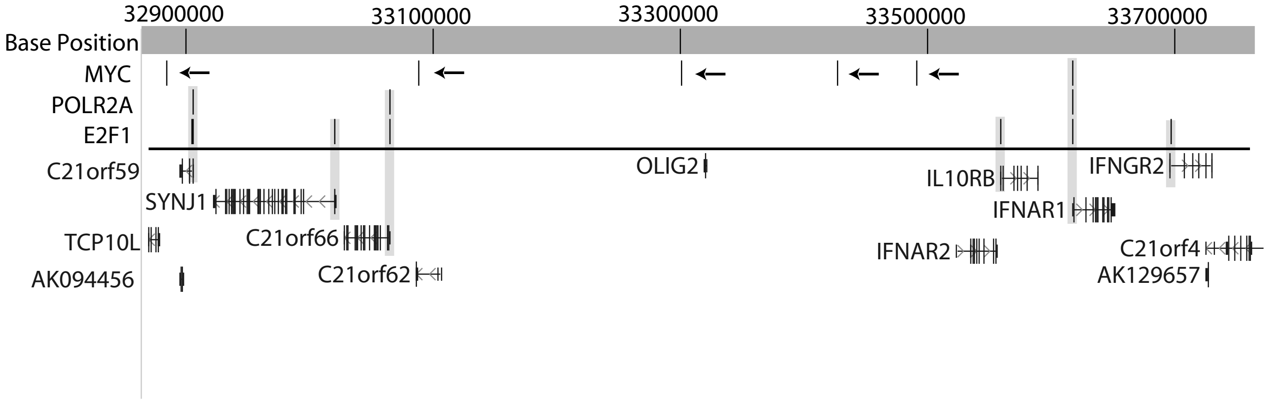
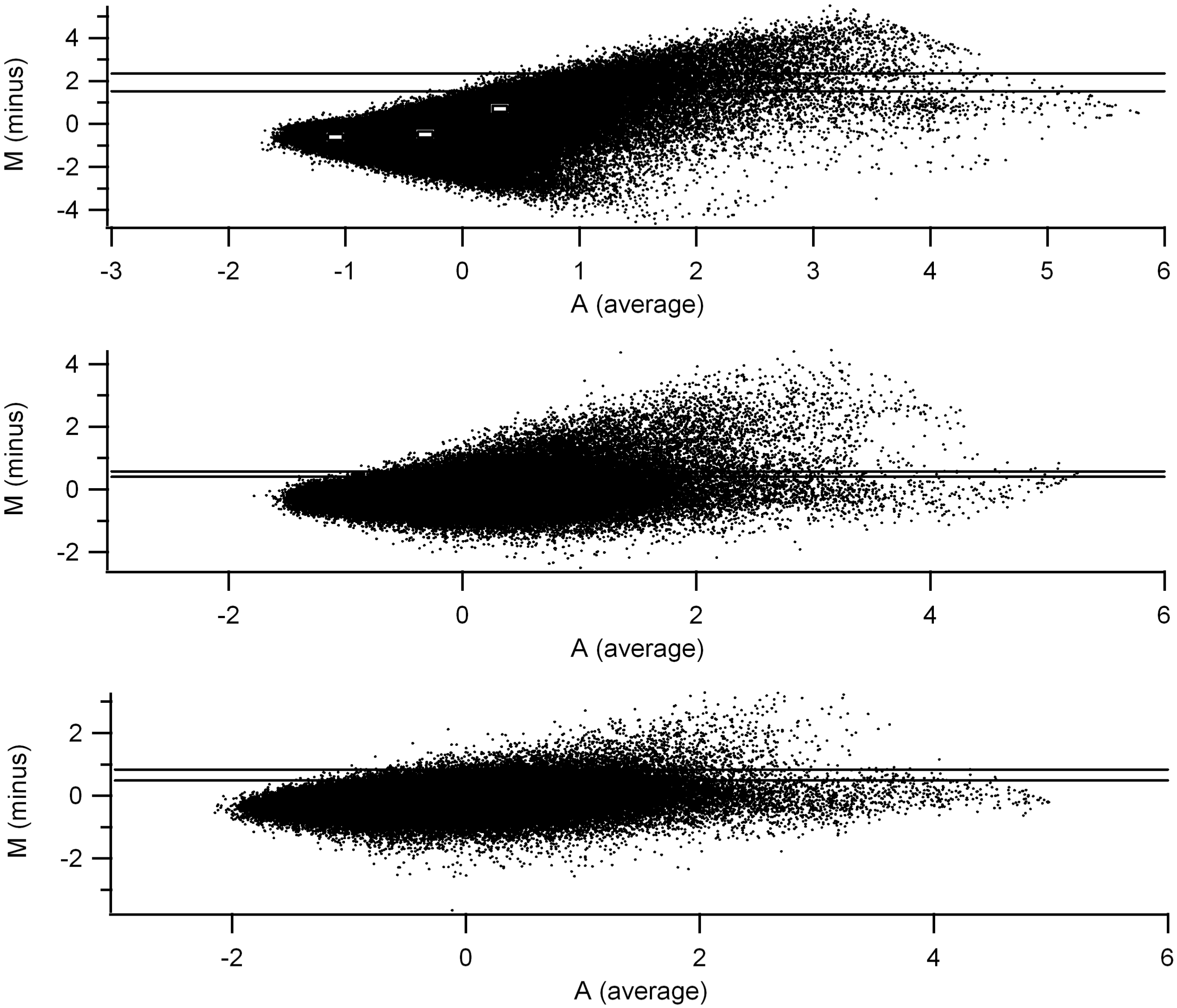

A
34700000
34800000
34900000
35000000
Base Position
E2F1_L1
CpG Islands
TTTSSCGC
TTTm1SSCGC
B
E2F1 BSs
25
180
- TTTSSCGC
+ TTTSSCGC
TTTSSCGC motifs
27
484
- E2F1 TFBS
+ E2F1 TFBS
C
E2F1 BSs
37
168
- CpG island
+ CpG island
CpG Islands
137
365
- E2F1 TFBS
+ E2F1 TFBS
D
bits
2
1
0
5′
1
2
3
4
5
6
7
8
3′
weblogo.berkeley.edu

Figure S5


32900000
33100000
33300000
33500000
33700000
Base Position
MYC
POLR2A
E2F1
C21orf59
SYNJ1
TCP10L
AK094456
C21orf66
C21orf62
OLIG2
IL10RB
IFNAR1
IFNAR2
IFNGR2
C21orf4
AK129657

Supplemental Figure S6



**Supplemental Figure S6. MA plots of HeLa E2F1 array data.
See Legend.**

Supplementary Table 1: PCR Confirmations of Peak Predictions

	total	# incorrect	% incorrect	"confirmation" rate
Predicted Sites				
Positives				
L1	29	0	0.0%	100.0%
first appears at L2	6	0	0.0%	100.0%
first appears at L3	3	1	33.3%	66.7%
first appears at L4	5	4	80.0%	20.0%
Negatives	10	1	10.0%	90.0%
Individual Array				
Positives				
L1	82	4	4.9%	95.1%
first appears at L2	13	0	0.0%	100.0%
first appears at L3	7	1	14.3%	85.7%
first appears at L4	11	5	45.5%	54.5%
Negatives	36	8	22.2%	77.8%

Supplementary Table 2: Distribution of E2F1 and myc putative TFBS by encode region

E2F1					**Myc**			
Region	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**
ENm001	7	11	10	12	4	14	13	33
ENm002	11	11	15	18	15	21	25	42
ENm003	2	2	2	3	4	6	7	18
ENm004	12	13	14	22	13	19	22	42
ENm005	19	17	24	27	14	28	27	62
ENm006	25	27	38	69	9	23	24	52
ENm007	15	18	21	32	4	6	11	25
ENm008	6	19	19	44	2	3	5	13
ENm009	0	2	1	5	4	10	13	15
ENm010	13	16	19	30	6	16	11	30
ENm011	1	3	6	18	1	4	6	11
ENm012	1	1	1	3	2	5	4	10
ENm013	5	7	8	9	5	5	6	12
ENm014	3	2	5	9	3	4	6	8
ENr111	2	3	4	7	1	3	3	10
ENr112	0	0	0	0	0	1	0	2
ENr113	0	0	0	1	0	0	0	0
ENr114	0	0	0	1	0	0	0	0
ENr121	3	4	4	7	4	6	6	12
ENr122	1	1	2	2	0	2	3	9
ENr123	1	5	4	6	0	3	1	4
ENr131	1	1	2	3	1	4	5	13
ENr132	4	12	12	41	5	12	11	25
ENr133	3	5	9	12	4	8	7	17
ENr211	0	1	0	2	6	9	12	19
ENr212	1	3	2	4	2	7	10	22
ENr213	0	0	0	1	0	0	0	1
ENr221	3	4	5	5	6	9	10	18
ENr222	0	0	0	0	3	4	3	7
ENr223	5	5	9	9	5	8	10	13
ENr231	9	10	14	20	6	10	9	18
ENr232	11	12	18	28	3	6	5	15
ENr233	8	10	9	10	7	16	11	30
ENr311	0	0	0	1	2	2	2	7
ENr312	0	0	0	4	9	14	16	22
ENr313	0	1	0	2	1	1	2	2
ENr321	1	2	2	2	1	4	3	14
ENr322	1	2	5	9	3	8	9	23
ENr323	4	6	10	10	3	5	4	9
ENr324	5	5	7	7	2	3	2	4
ENr331	2	5	6	13	3	4	8	12
ENr332	8	13	14	35	3	5	9	18
ENr333	8	9	10	17	3	7	8	14
ENr334	4	7	6	15	3	7	5	17
Total-ENm	120	149	183	301	86	164	180	373
Total-ENr	85	126	154	274	86	168	174	377
Total-all	205	275	337	575	172	332	354	750
Regulatory Regions	170	237	253	460	ND	ND	ND	ND

Table S3: Novel transcripts with high stringency E2F1 BS within 1 kb

Gencode class	transcriptname
TEC	AC002064.7
Putative_gencode_conf	AC002456.2
Novel_transcript	AC004080.1
Putative	AC004080.13
Novel_transcript	AC008937.3
TEC	AC009303.2
Novel_transcript	AC012314.6
Novel_transcript_gencode_conf	AC018512.2
Novel_transcript	AC034220.1
Novel_transcript_gencode_conf	AC034220.3
Putative_gencode_conf	AC053503.9
Novel_transcript	AC106873.1
TEC	AP000269.2
Novel_transcript_gencode_conf	AP000279.69
Novel_transcript_gencode_conf	AP000295.7
Putative	RP1-315G1.3
Putative_gencode_conf	RP11-126K1.2
Putative	RP11-126K1.6
Novel_transcript	RP11-223E19.2
Novel_transcript	RP11-247A12.2
Novel_transcript_gencode_conf	RP11-328M4.2
Novel_transcript_gencode_conf	RP11-398K22.12
Novel_transcript_gencode_conf	RP11-65J3.1
Putative	RP4-614O4.8
Novel_transcript	RP5-931E15.2
Processed_pseudogene	RP5-931E15.3
TEC	XX-FW81657B9.5
TEC	XX-FW83563B9.5

Supplementary Table S4:
E2F1 binding site distance to nearest
5' transcription start site and 3' transcription end site

Distance to nearest 5' TSS

Distance	L1	L2	L3
Zero	50.7%	48.7%	38.9%
<1Kb	81.5%	75.3%	74.2%
<2Kb	84.9%	80.0%	79.5%
<5kb	89.8%	85.5%	86.1%
<10kb	92.2%	90.9%	90.5%
>20kb	4.4%	5.8%	6.5%

Distance to nearest 3' TES

Distance	L1	L2	L3
Zero	3.9%	7.6%	3.6%
<1Kb	19.0%	23.3%	18.4%
<2Kb	32.7%	33.1%	32.0%
<5kb	47.3%	50.9%	46.9%
<10kb	60.5%	63.6%	59.6%
>20kb	21.5%	22.9%	25.2%