nature | methods

Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data

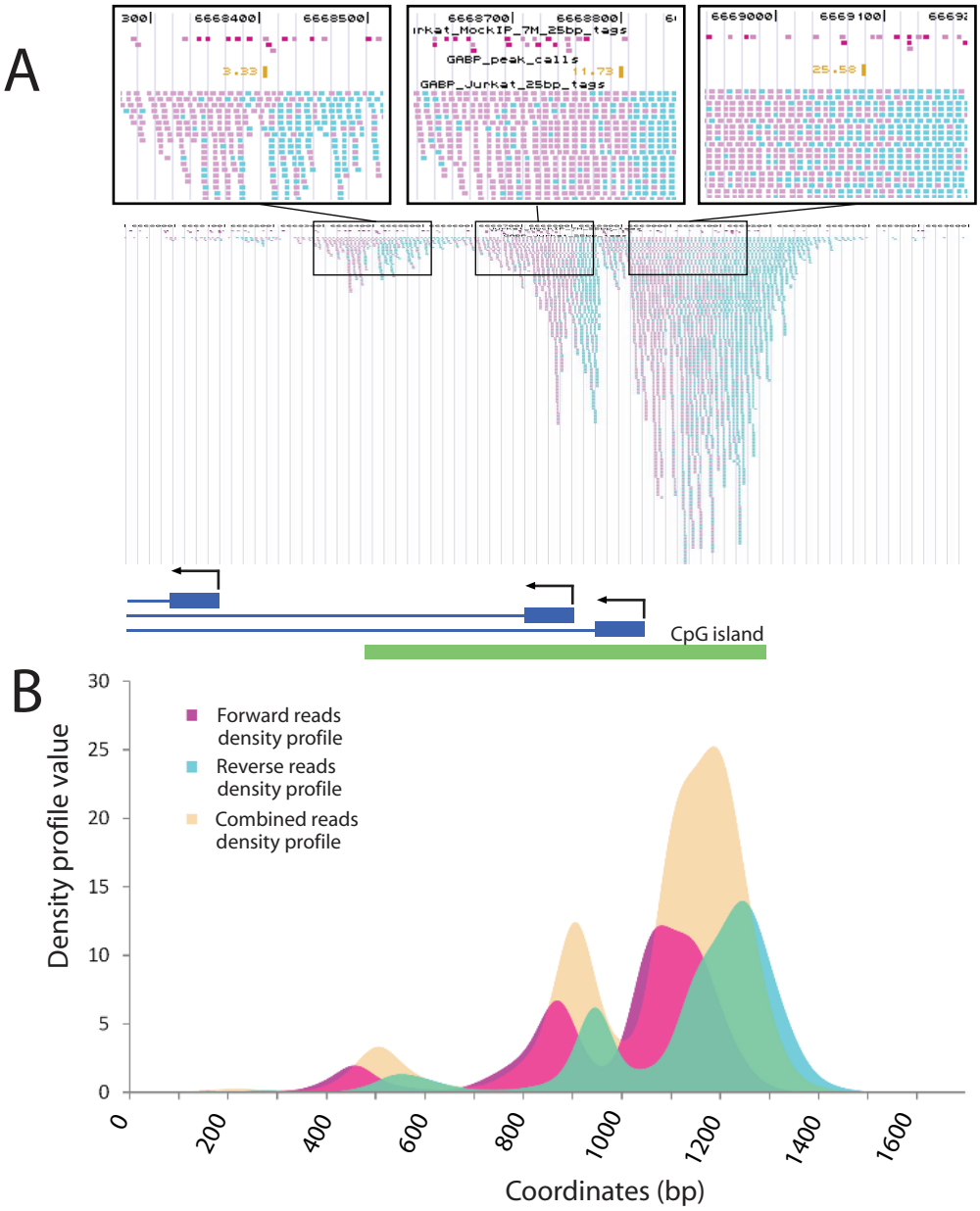
Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton,

Serafim Batzoglou, Richard M Myers & Arend Sidow

Supplementary figures and text:

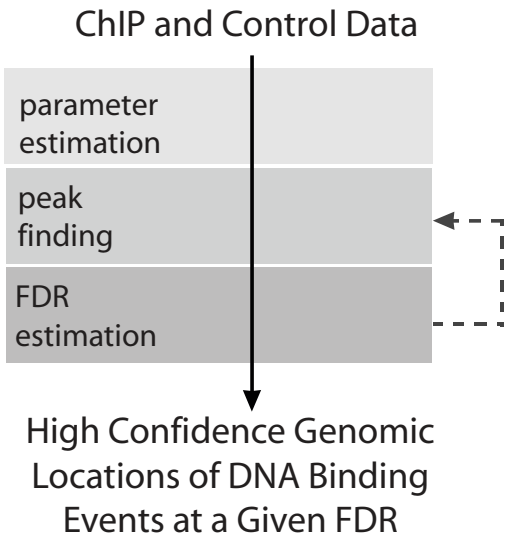
Supplementary Figure 1	Resolution of closely spaced peaks by QuEST.
Supplementary Figure 2	QuEST analysis outline.
Supplementary Figure 3	Peak saturation curves at 1% FDR.
Supplementary Figure 4	MAST Curves.
Supplementary Figure 5	Expression analysis of peak-associated genes.
Supplementary Figure 6	Distribution of peak distances between forward and reverse KDE profiles.
Supplementary Table 1	GABP GO categories.
Supplementary Table 2	SRF GO categories.
Supplementary Table 3	NRSF GO categories.
Supplementary Methods	

Supplementary Figure 1. Resolution of closely spaced peaks by QuEST.



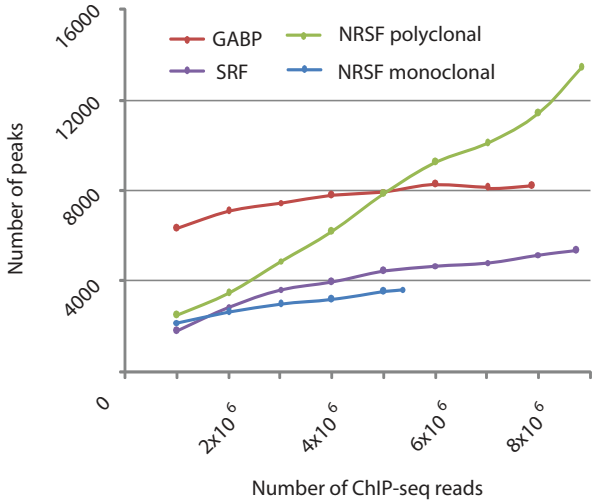
(A) GABP ChIP-Seq reads, RX-noIP reads and peak calls at the promoter of Nuclear Matrix Transcription Factor 4 gene (ZNF384). GABP ChIP forward reads (purple) and reverse reads (blue) are displayed as small bands 5 bp in width. Three portions of the browser are magnified to show the relatively sparse number of reads in the RX-noIP track (top) and the GABP peak calling track with the peak score values and positions (displayed as small yellow bands). Shown below the sequencing data are three first exons of distinct human transcripts that originate in the region. **(B)** The forward (purple), reverse (blue), and combined (yellow) density profiles derived from these data illustrate that the three peaks are cleanly separated by QuEST, and that the heights of the peaks quantify enrichment of the ChIP tags. The zero coordinate of the graph corresponds to position 6667900 of Chromosome 12, NCBI build 36.

Supplementary Figure 2. QuEST analysis outline.



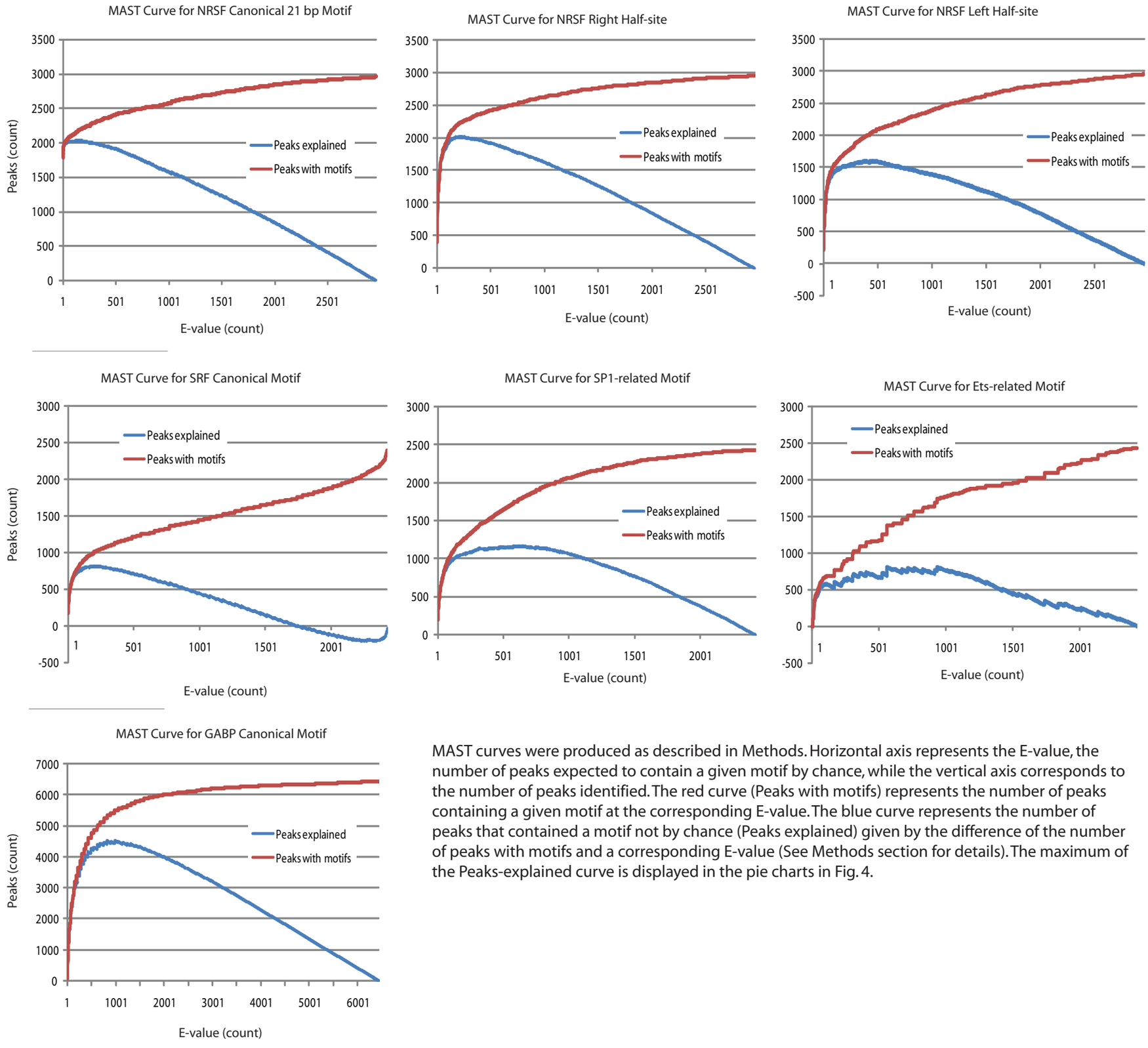
High-throughput sequencing data from ChIP and RX-noIP libraries are used to first estimate experiment-specific statistical parameters and then to perform peak calling. ChIP peak calls and pseudo-ChIP peak calls together provide an estimate of the false discovery rate. Peak calling can be repeated with varying thresholds until the desired FDR is achieved (dashed arrow).

Supplementary Figure 3. Peak saturation curves at 1% FDR.



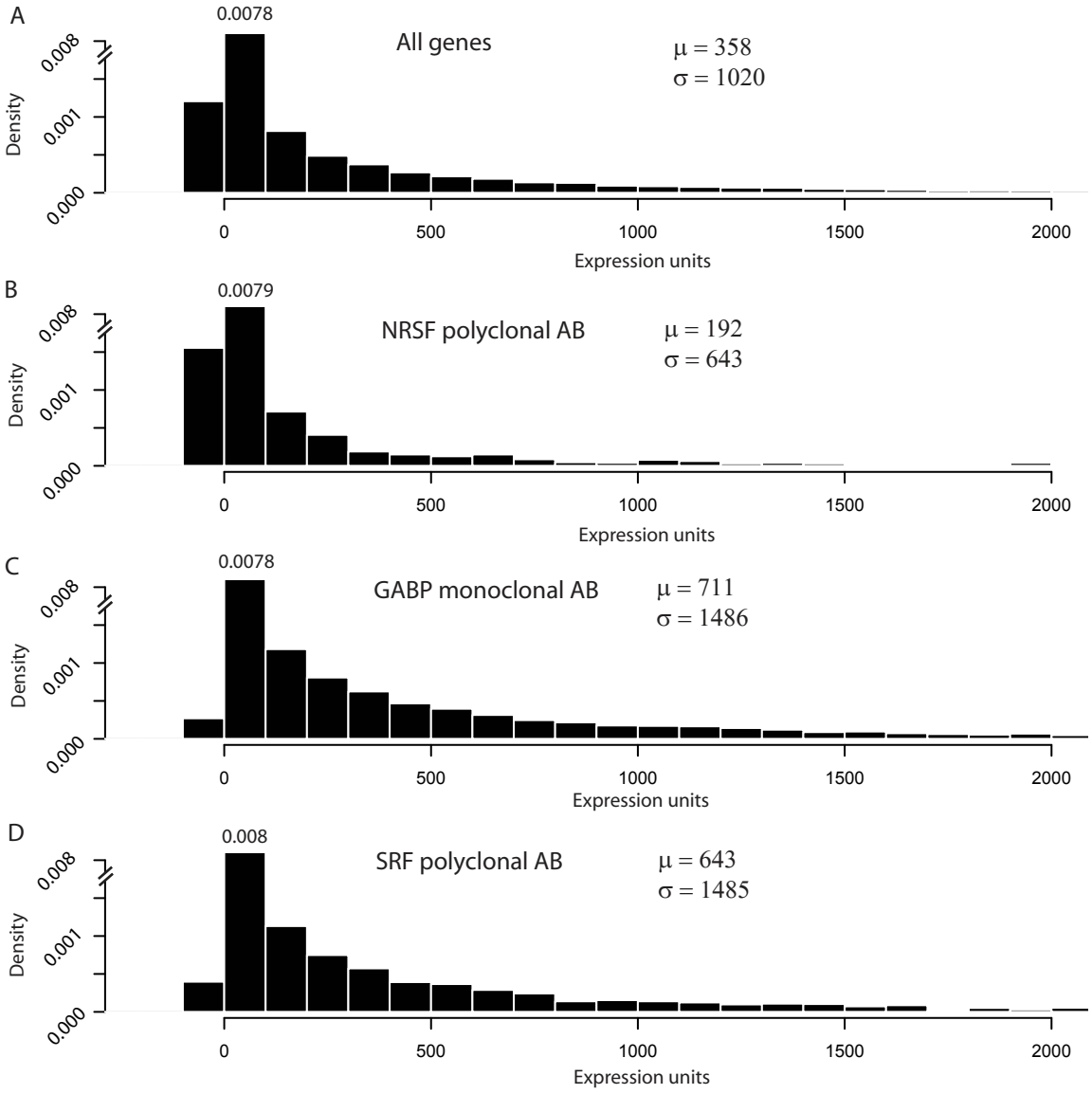
For each of the four datasets, the number of peak calls as a function of the number of ChIP-Seq reads that were used to perform peak finding, is plotted (See Supplemental Methods for details).

Supplementary Figure 4. MAST curves.



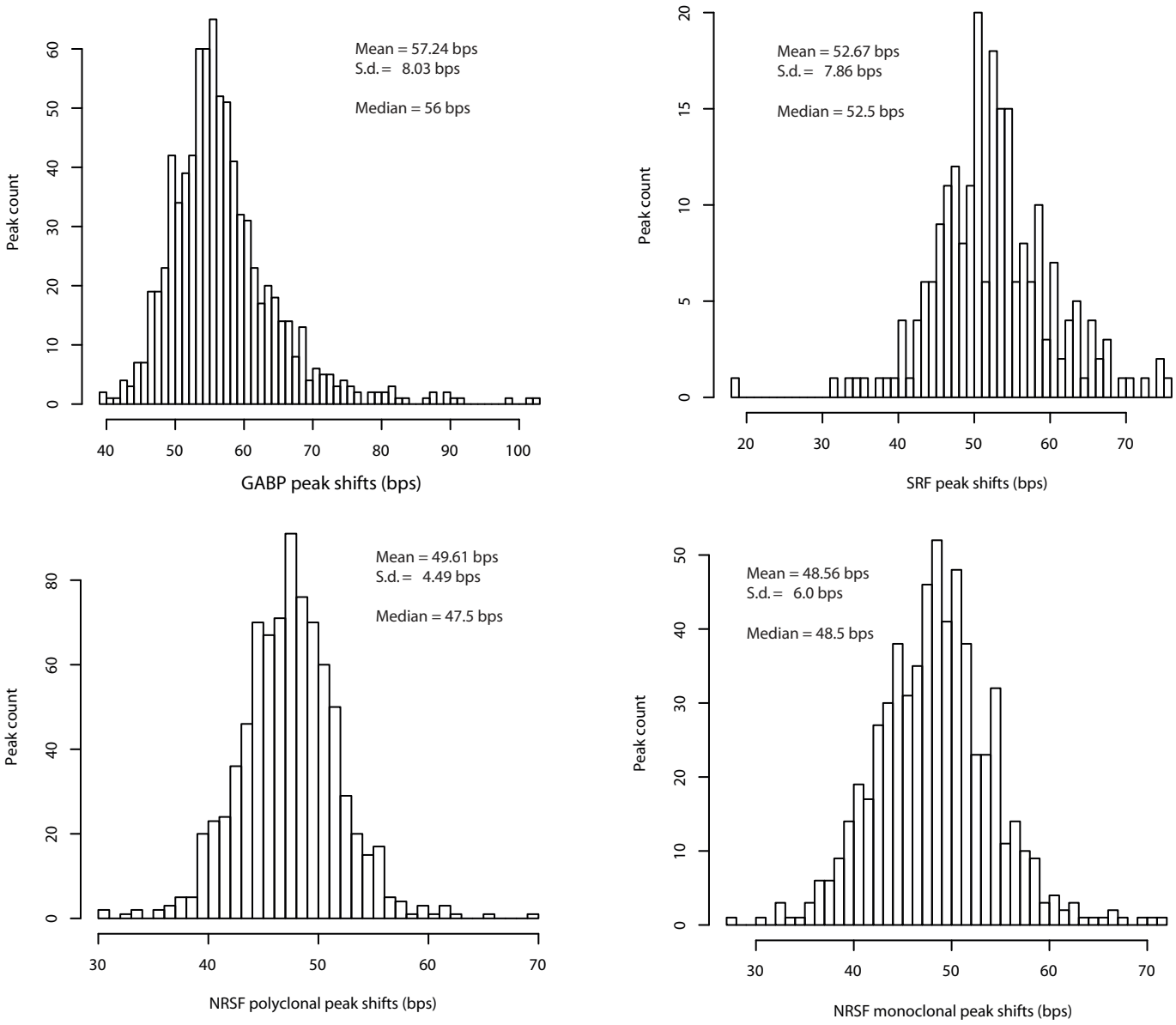
MAST curves were produced as described in Methods. Horizontal axis represents the E-value, the number of peaks expected to contain a given motif by chance, while the vertical axis corresponds to the number of peaks identified. The red curve (Peaks with motifs) represents the number of peaks containing a given motif at the corresponding E-value. The blue curve represents the number of peaks that contained a motif not by chance (Peaks explained) given by the difference of the number of peaks with motifs and a corresponding E-value (See Methods section for details). The maximum of the Peaks-explained curve is displayed in the pie charts in Fig. 4.

Supplementary Figure 5. Expression analysis of peak-associated genes.



Panels display distributions of expression values from an Illumina expression BeadArray, and their mean and standard deviations. (A), distribution of all genes assayed on the array. (B) - (D), distributions of genes associated with the indicated factors.

Supplementary Figure 6. Distribution of peak distances between forward and reverse KDE profiles.



Peak distances identified from a core set of enriched regions (described in Methods, Peak Shift Estimate section) are displayed as histograms for each dataset. The peak shift estimate (either mean or median) is robust across all 4 of analyzed ChIP datasets and is highly concordant between two NRSF experiments.

Supplementary Table 1. GABP GO categories.

GO group	Instances	p-value		Annotation
GO:0005634	1493 out of 4085	1.21E-240	enriched	nucleus
GO:0005515	1239 out of 4063	1.87E-120	enriched	protein binding
GO:0006350	517 out of 1370	2.03E-80	enriched	transcription
GO:0000166	587 out of 1671	1.13E-77	enriched	nucleotide binding
GO:0046872	691 out of 2155	9.06E-73	enriched	metal ion binding
GO:0003723	266 out of 543	4.19E-68	enriched	RNA binding
GO:0005622	681 out of 2169	1.60E-67	enriched	intracellular
GO:0003677	439 out of 1184	2.35E-65	enriched	DNA binding
GO:0008270	688 out of 2240	5.12E-64	enriched	zinc ion binding
GO:0005739	320 out of 759	2.98E-62	enriched	mitochondrion
GO:0005737	525 out of 1564	1.02E-61	enriched	cytoplasm
GO:0008380	130 out of 196	9.15E-54	enriched	RNA splicing
GO:0005524	446 out of 1352	6.41E-50	enriched	ATP binding
GO:0006355	538 out of 1818	3.17E-44	enriched	regulation of transcription, DNA-dependent
GO:0005681	86 out of 118	4.45E-41	enriched	spliceosome complex
GO:0016740	368 out of 1120	9.85E-41	enriched	transferase activity
GO:0006397	106 out of 178	9.91E-38	enriched	mRNA processing
GO:0015031	148 out of 314	6.86E-36	enriched	protein transport
GO:0007186	30 out of 817	4.71E-33	depleted	G-protein coupled receptor protein signaling pathway
GO:0006512	154 out of 356	5.68E-32	enriched	ubiquitin cycle
GO:0005783	221 out of 638	2.60E-28	enriched	endoplasmic reticulum
GO:0006281	84 out of 148	3.85E-28	enriched	DNA repair
GO:0016787	270 out of 864	2.71E-26	enriched	hydrolase activity
GO:0007608	4 out of 392	3.26E-26	depleted	sensory perception of smell
GO:0004872	106 out of 1407	3.49E-26	depleted	receptor activity
GO:0007049	165 out of 439	8.37E-26	enriched	cell cycle
GO:0006457	112 out of 247	9.78E-26	enriched	protein folding
GO:0005730	67 out of 115	1.23E-23	enriched	nucleolus
GO:0030529	62 out of 108	1.60E-21	enriched	ribonucleoprotein complex
GO:0003676	237 out of 786	5.23E-21	enriched	nucleic acid binding
GO:0050896	23 out of 550	1.10E-20	depleted	response to stimulus
GO:0005576	36 out of 681	1.11E-20	depleted	extracellular region
GO:0006412	136 out of 391	3.61E-18	enriched	protein biosynthesis
GO:0016874	97 out of 242	5.71E-18	enriched	ligase activity
GO:0005842	31 out of 40	5.52E-17	enriched	cytosolic large ribosomal subunit (sensu Eukaryota)
GO:0005654	46 out of 79	1.10E-16	enriched	nucleoplasm
GO:0006888	43 out of 71	1.43E-16	enriched	ER to Golgi vesicle-mediated transport
GO:0006364	34 out of 49	5.01E-16	enriched	rRNA processing

GO:0003674	184 out of 621	7.69E-16	enriched	molecular_function
GO:0006260	55 out of 112	2.88E-15	enriched	DNA replication
GO:0005843	27 out of 35	7.15E-15	enriched	cytosolic small ribosomal subunit (sensu Eukaryota)
GO:0006464	88 out of 242	1.46E-13	enriched	protein modification
GO:0051082	52 out of 113	4.16E-13	enriched	unfolded protein binding
GO:0003684	27 out of 39	5.68E-13	enriched	damaged DNA binding
GO:0006886	68 out of 171	6.78E-13	enriched	intracellular protein transport
GO:0003735	110 out of 340	1.09E-12	enriched	structural constituent of ribosome
GO:0016301	74 out of 196	1.42E-12	enriched	kinase activity
GO:0051301	68 out of 174	1.75E-12	enriched	cell division
GO:0000398	25 out of 36	3.71E-12	enriched	nuclear mRNA splicing, via spliceosome
GO:0006915	104 out of 321	3.94E-12	enriched	apoptosis
GO:0006511	58 out of 140	4.34E-12	enriched	ubiquitin-dependent protein catabolism
GO:0006468	151 out of 528	5.20E-12	enriched	protein amino acid phosphorylation
GO:0001584	11 out of 280	1.12E-11	depleted	rhodopsin-like receptor activity
GO:0006310	28 out of 46	2.28E-11	enriched	DNA recombination
GO:0005488	110 out of 356	2.53E-11	enriched	binding
GO:0008168	43 out of 93	3.42E-11	enriched	methyltransferase activity
GO:0005793	22 out of 31	3.77E-11	enriched	ER-Golgi intermediate compartment
GO:0004674	111 out of 363	4.13E-11	enriched	protein serine/threonine kinase activity
GO:0008150	173 out of 645	4.79E-11	enriched	biological_process
GO:0008033	27 out of 45	8.26E-11	enriched	tRNA processing
GO:0005525	114 out of 381	9.77E-11	enriched	GTP binding
GO:0005829	101 out of 325	1.09E-10	enriched	cytosol
GO:0042802	62 out of 166	1.40E-10	enriched	protein self binding
GO:0003899	24 out of 38	2.04E-10	enriched	DNA-directed RNA polymerase activity
GO:0017111	38 out of 81	2.70E-10	enriched	nucleoside-triphosphatase activity
GO:0005743	54 out of 138	2.84E-10	enriched	mitochondrial inner membrane
GO:0004842	51 out of 128	4.12E-10	enriched	ubiquitin-protein ligase activity
GO:0031202	17 out of 22	7.60E-10	enriched	RNA splicing factor activity, transesterification mechanism
GO:0003713	58 out of 157	8.48E-10	enriched	transcription coactivator activity
GO:0003743	32 out of 65	1.48E-09	enriched	translation initiation factor activity
GO:0005840	85 out of 271	1.78E-09	enriched	ribosome
GO:0005643	29 out of 57	3.25E-09	enriched	nuclear pore
GO:0000502	15 out of 19	4.74E-09	enriched	proteasome complex (sensu Eukaryota)
GO:0016853	44 out of 110	5.28E-09	enriched	isomerase activity
GO:0005575	183 out of 737	7.82E-09	enriched	cellular_component
GO:0008026	30 out of 62	7.84E-09	enriched	ATP-dependent helicase activity
GO:0016568	43 out of 108	9.15E-09	enriched	chromatin modification
GO:0000151	26 out of 50	1.15E-08	enriched	ubiquitin ligase complex
GO:0005789	29 out of 60	1.42E-08	enriched	endoplasmic reticulum membrane

GO:0003924	66 out of 201	1.45E-08	enriched	GTPase activity
GO:0006366	62 out of 186	2.06E-08	enriched	transcription from RNA polymerase II promoter
GO:0003697	22 out of 39	2.18E-08	enriched	single-stranded DNA binding
GO:0007165	216 out of 1734	2.75E-08	depleted	signal transduction
GO:0016567	24 out of 46	3.81E-08	enriched	protein ubiquitination
GO:0007264	67 out of 210	3.90E-08	enriched	small GTPase mediated signal transduction
GO:0000074	71 out of 230	6.73E-08	enriched	regulation of progression through cell cycle
GO:0008565	30 out of 68	1.03E-07	enriched	protein transporter activity
GO:0007067	47 out of 133	1.37E-07	enriched	mitosis
GO:0005615	41 out of 467	1.45E-07	depleted	extracellular space
GO:0006396	25 out of 52	1.54E-07	enriched	RNA processing
GO:0005096	48 out of 138	1.74E-07	enriched	GTPase activator activity
GO:0005777	34 out of 84	1.87E-07	enriched	peroxisome
GO:0004004	14 out of 20	1.94E-07	enriched	ATP-dependent RNA helicase activity
GO:0006506	14 out of 20	1.94E-07	enriched	GPI anchor biosynthesis
GO:0006461	42 out of 115	2.19E-07	enriched	protein complex assembly
GO:0003954	20 out of 37	2.40E-07	enriched	NADH dehydrogenase activity
GO:0004519	25 out of 53	2.43E-07	enriched	endonuclease activity
GO:0000781	12 out of 16	4.64E-07	enriched	chromosome, telomeric region
GO:0016251	14 out of 21	4.85E-07	enriched	general RNA polymerase II transcription factor activity
GO:0003700	212 out of 933	6.58E-07	enriched	transcription factor activity
GO:0048471	24 out of 52	6.81E-07	enriched	perinuclear region
GO:0005794	59 out of 191	7.37E-07	enriched	Golgi apparatus
GO:0019901	23 out of 49	8.00E-07	enriched	protein kinase binding
GO:0016564	33 out of 86	1.13E-06	enriched	transcriptional repressor activity
GO:0000287	93 out of 350	1.42E-06	enriched	magnesium ion binding
GO:0006446	17 out of 31	1.45E-06	enriched	regulation of translational initiation

Supplementary Table 2. SRF GO categories.

GO group	Instances	p-value		Annotation
GO:0005634	515 out of 4085	5.18E-70	enriched	nucleus
GO:0005515	467 out of 4063	2.16E-50	enriched	protein binding
GO:0005737	202 out of 1564	6.67E-27	enriched	cytoplasm
GO:0006350	183 out of 1370	4.13E-26	enriched	transcription
GO:0000166	203 out of 1671	1.04E-23	enriched	nucleotide binding
GO:0046872	239 out of 2155	1.92E-22	enriched	metal ion binding
GO:0008270	238 out of 2240	5.76E-20	enriched	zinc ion binding
GO:0003723	88 out of 543	3.38E-18	enriched	RNA binding
GO:0006355	191 out of 1818	1.15E-15	enriched	regulation of transcription, DNA-dependent
GO:0005524	152 out of 1352	4.84E-15	enriched	ATP binding
GO:0005622	211 out of 2169	8.63E-14	enriched	intracellular
GO:0005739	96 out of 759	7.94E-13	enriched	mitochondrion
GO:0016740	125 out of 1120	2.17E-12	enriched	transferase activity
GO:0003677	130 out of 1184	2.44E-12	enriched	DNA binding
GO:0015031	51 out of 314	3.18E-11	enriched	protein transport
GO:0007186	12 out of 817	1.15E-10	depleted	G-protein coupled receptor protein signaling pathway
GO:0005681	28 out of 118	1.43E-10	enriched	spliceosome complex
GO:0006397	35 out of 178	2.04E-10	enriched	mRNA processing
GO:0008380	36 out of 196	7.95E-10	enriched	RNA splicing
GO:0003700	101 out of 933	1.34E-09	enriched	transcription factor activity
GO:0006281	30 out of 148	1.80E-09	enriched	DNA repair
GO:0005783	74 out of 638	1.21E-08	enriched	endoplasmic reticulum
GO:0006928	26 out of 127	1.67E-08	enriched	cell motility
GO:0006468	64 out of 528	2.20E-08	enriched	protein amino acid phosphorylation
GO:0007049	56 out of 439	2.86E-08	enriched	cell cycle
GO:0005856	47 out of 340	3.20E-08	enriched	cytoskeleton
GO:0016874	36 out of 242	1.99E-07	enriched	ligase activity
GO:0003779	36 out of 244	2.42E-07	enriched	actin binding
GO:0007265	12 out of 34	2.47E-07	enriched	Ras protein signal transduction
GO:0050896	9 out of 550	4.60E-07	depleted	response to stimulus
GO:0004674	46 out of 363	5.27E-07	enriched	protein serine/threonine kinase activity
GO:0008134	18 out of 82	8.42E-07	enriched	transcription factor binding
GO:0016568	21 out of 108	8.86E-07	enriched	chromatin modification
GO:0003714	21 out of 110	1.20E-06	enriched	transcription corepressor activity
GO:0030036	19 out of 94	1.56E-06	enriched	actin cytoskeleton organization and biogenesis
GO:0005576	15 out of 681	1.64E-06	depleted	extracellular region

Supplementary Table 3. NRSF GO categories.

GO group	Instances	p-value		Annotation
GO:0016020	364 out of 4640	1.33E-43	enriched	membrane
GO:0006811	89 out of 461	9.35E-37	enriched	ion transport
GO:0005509	117 out of 897	4.27E-31	enriched	calcium ion binding
GO:0007268	50 out of 178	4.45E-29	enriched	synaptic transmission
GO:0016021	265 out of 3481	2.67E-28	enriched	integral to membrane
GO:0005216	42 out of 183	5.82E-21	enriched	ion channel activity
GO:0007399	53 out of 302	2.19E-20	enriched	nervous system development
GO:0030955	33 out of 116	7.49E-20	enriched	potassium ion binding
GO:0006813	37 out of 153	1.87E-19	enriched	potassium ion transport
GO:0005515	265 out of 4063	1.14E-18	enriched	protein binding
GO:0045202	25 out of 80	1.85E-16	enriched	synapse
GO:0005887	98 out of 1048	5.79E-16	enriched	integral to plasma membrane
GO:0005886	66 out of 652	1.08E-12	enriched	plasma membrane
GO:0007155	56 out of 514	3.16E-12	enriched	cell adhesion
GO:0045211	21 out of 89	1.87E-11	enriched	postsynaptic membrane
GO:0005249	19 out of 72	2.07E-11	enriched	voltage-gated potassium channel activity
GO:0007411	16 out of 53	8.62E-11	enriched	axon guidance
GO:0006816	19 out of 88	8.38E-10	enriched	calcium ion transport
GO:0005245	10 out of 22	3.07E-09	enriched	voltage-gated calcium channel activity
GO:0008076	18 out of 87	4.71E-09	enriched	voltage-gated potassium channel complex
GO:0043025	11 out of 29	5.01E-09	enriched	cell soma
GO:0000166	112 out of 1671	6.06E-09	enriched	nucleotide binding
GO:0007165	115 out of 1734	6.72E-09	enriched	signal transduction
GO:0005234	9 out of 19	1.23E-08	enriched	glutamate-gated ion channel activity
GO:0008021	13 out of 48	2.13E-08	enriched	synaptic vesicle
GO:0005737	104 out of 1564	3.05E-08	enriched	cytoplasm
GO:0007269	9 out of 21	3.62E-08	enriched	neurotransmitter secretion
GO:0006814	19 out of 111	4.45E-08	enriched	sodium ion transport
GO:0005891	9 out of 22	5.88E-08	enriched	voltage-gated calcium channel complex
GO:0004872	94 out of 1407	1.10E-07	enriched	receptor activity
GO:0005524	91 out of 1352	1.27E-07	enriched	ATP binding
GO:0005215	34 out of 326	1.39E-07	enriched	transporter activity
GO:0031402	16 out of 88	2.13E-07	enriched	sodium ion binding
GO:0005624	43 out of 484	3.07E-07	enriched	membrane fraction
GO:0005244	10 out of 37	9.17E-07	enriched	voltage-gated ion channel activity
GO:0019992	12 out of 56	1.12E-06	enriched	diacylglycerol binding
GO:0030424	10 out of 38	1.20E-06	enriched	axon
GO:0030425	8 out of 23	1.39E-06	enriched	dendrite

GO:0005085	17 out of 113	1.40E-06	enriched	guanyl-nucleotide exchange factor activity
GO:0006836	10 out of 39	1.55E-06	enriched	neurotransmitter transport

Supplemental Methods

Density Profiles (1) Density profiles were defined as in Methods using the following formula: $H_{+/-}(i) = \frac{1}{h} \sum_{j=i-3h}^{i+3h} K\left((j-i)/h\right) \cdot C_{+/-}(j)$. The sum in the expression for H is limited to points within 3 times the KDE bandwidth, h, to improve computational performance. Points that are further away only contribute one percent or less compared to sample points occurring at positions where H is evaluated. We used $h=30$ in all analyses. This choice was adopted after close inspection of a collection of enriched regions, in which optimal bandwidth was estimated using the statistical package R. Estimates varied between 6 to 9 bps depending on the region, but close visual inspection of the resulting density estimates at these loci revealed multiple sub-peaks within each region. These are likely due to biases in the ChIP, chromatin shearing, PCR or sequencing rather than biological phenomena. We therefore adjusted the bandwidth to a higher value until the profiles at these enriched regions appeared sufficiently smooth. We note that our analyses of the ChIP-Seq data in the paper did not apparently suffer in resolution and accuracy by this conservative adjustment of the KDE bandwidth.

Two-step Motif Analysis. As described in the manuscript, we ran the peak-associated sequences through MEME to discover motifs. Although MEME implements probabilistic subtraction allowing discovery of multiple motifs in the data, cofactor motifs may not be statistically significant unless a subset of data consisting of sequences free of canonical motifs is considered. We therefore conducted a second pass of MEME through the sequences that contained no strong evidence of a canonical motif, excluding sequences that contained motifs at a log-odds score of 3.0 (corresponding to uncorrected P-values between 0.0001 and 0.001) and above. We found this two-step strategy to be effective, as evidenced by the discovery of the Ets motif in the SRF dataset.

Number of Sequence Reads Required to Achieve Peak Number Saturation. A critical question for ChIP-Seq experiments in general and our datasets in particular is whether the number of reads produced in the experiment reached saturation, or whether more reads would have substantially increased the number of called peaks. To evaluate whether we sequenced our libraries to saturation, we performed QuEST analyses to identify significant peaks for each library on data subsets, which varied in size from 1 million reads to the total number in the experiment, in increments of 1 million. The background tag count was fixed at 7 million (Fig. S3; Methods), and the FDR at 1%. This resulting peak saturation curves exhibit an up to 50% gain in the number of identified peaks going from 1M to 4M reads (SRF, NRSF polyclonal) with a visible flattening at around 5M reads (SRF, GABP). The number of peaks for the NRSF polyclonal data increased linearly across the entire range without achieving flattening. This may be due to unspecific interactions between the polyclonal antibody and other

proteins, and suggests that peak saturation is experiment-specific and ideally should be evaluated for each experiment.

Number of Peaks Explained by Motifs. To determine how many peaks could be explained by statistically significant motifs, we estimated how many sequences had motifs that could be explained by chance. We first analyzed the peak-associated sequences for motif occurrences using the MEME PSSMs by counting the number of motif-containing sequences at various stringencies with the MEME tool MAST (Bailey TL and Gribskov M, *Bioinformatics* **14**:48-54, 1998).

MAST reports the number of motif containing sequences in the input set of sequences at a particular stringency. The stringency is given by an E-value, which is colloquially defined as the number of sequences expected to contain the motif by chance. The formal definition is "an estimate of the number of motifs (with the same width and number of occurrences) that would have equal or higher log likelihood ratio if the training set sequences had been generated randomly according to the (0-order portion of the) background model[22]".

We first counted the number of motif-containing sequences across a range of E-values. As E-values increase, the number of predicted motifs increases (Supplementary Figure 4, red lines), due to increased sensitivity of the predictions. For any given E-value, we then calculated the fraction of peaks explained by that motif by subtracting the E-value from the total number of sequences containing the motif (at that same stringency). Applying this approach to a range of E-values, we obtained curves that plot the fraction of peaks explained by a motif at any given E-value (Supplementary Figure 4, blue lines). We then interpreted the maximum of this "MAST curve" as the maximum fraction of peaks that can be explained by that particular motif. This number is reported as a percentage of the total number of peaks in Fig. 4. In principle, this approach may yield an optimistic estimate of the explained peak fraction if the E-value estimate is noisy; we note, however, that MEME E-values have been demonstrated to be quite accurate[22] and therefore regard our estimate as a reasonably realistic approximation.

Peak Saturation Analysis. The peaks were identified using QuEST thresholds corresponding to 1% FDR. To do so, we used a rescue ratio of 10 and a background CDP threshold was chosen to correspond to 3 fold enrichment (i.e. if the ChIP CDP threshold is T, then the background CDP threshold is T times the number of background reads divided by three times the number of ChIP reads). Then, both ChIP CDP and background CDP thresholds were varied together until an FDR of 1% was achieved. Peaks were called using ChIP subsets of multiples of 1 million tags up to the full ChIP dataset. Background data contained 7M RX_noIP reads (Fig. S3). The FDR estimate was provided using the same number of pseudo ChIP reads as in the corresponding subset of the ChIP data.

Our data sets provide some guidance as to how many reads are required to capture a sufficiently comprehensive set of peaks so as to approach saturation. We estimate that for the human genome, 5 - 10 million reads, or approximately two to six lanes of a Solexa run (depending on library quality and other parameters), is a good starting point (Fig. S3), but an evaluation of peak saturation curves may be necessary to assess whether the library was sequenced to sufficient depth. The cost of such experiments seems well worth the high data quality that ChIP-Seq produces.

ChIP-Seq Library Construction and Sequencing. The Jurkat human T lymphoblast cell line was cultured according to standard protocols (ATCC). NRSF/REST chromatin immunoprecipitation (ChIP) was performed as previously[5] with a custom monoclonal antibody and with a polyclonal antibody (Upstate AB15548). SRF ChIP was performed as described previously[17] using polyclonal antibody from Santa Cruz Biotechnology (sc-335), and GABP ChIP was performed as described elsewhere[20] using a monoclonal antibody from Santa Cruz (sc28312). For each ChIP, we collected 10-20 million cells and then added 1% formaldehyde to crosslink the proteins to the DNA. We then lysed and collected the nuclei and sonicated the chromatin to a final size of 500-1000 base pairs. We then coupled 5μL of each antibody to sheep anti-mouse magnetic beads (Invitrogen) by overnight incubation at 4°C in 5mg/ml BSA. We then added chromatin to the antibody-coupled beads and precipitated DNA-protein complexes overnight at 4°C. Four chromatin immunoprecipitations, corresponding to one batch of chromatin, were pooled for each library preparation. The control library input was chromatin that was reverse crosslinked, phenol extracted, and purified on a QIAQuick PCR cleanup column (Qiagen). The libraries were prepared as per Illumina's instructions (www.illumina.com), with some modifications. Briefly, the four ChIPs were blunted, phosphorylated, and then ligated to library adapters. We modified the protocol to include a PCR preamplification (30 sec at 98^{o}C; [10 sec at 98^{o}C, 30 sec at 65^{o}C, 30 sec at 72^{o}C] x 25 cycles; 5 min at 72^{o}C) following adapter ligation. After this PCR amplification, size selection was performed by gel electrophoresis and subsequent excision and purification of DNA in the ~150-300bp range. Following gel extraction, a second PCR amplification was performed (30 sec at 98^{o}C; [10 sec at 98^{o}C, 30 sec at 65^{o}C, 30 sec at 72^{o}C] x 18 cycles; 5 min at 72^{o}C). The control library was prepared in an identical manner, using ~50ng of input control DNA. These products were then sequenced on the Solexa 1G Genome Analyzer at 2-3pM concentration.

Gene Expression. Total RNA was prepared from three separate growths of Jurkat cells with Trizol (Invitrogen) according to manufacturer's protocols. Purity of the RNA was assessed by NanoDrop. RNA was amplified using standard procedures from the Illumina TotalPrep RNA amplification kit. We then used Illumina HumanRef-8 v2 Expression BeadChips for whole genome expression analysis on each of the biological replicates. Hybridization was performed using standard operating procedures from Illumina.

References are numbered according to the main text.