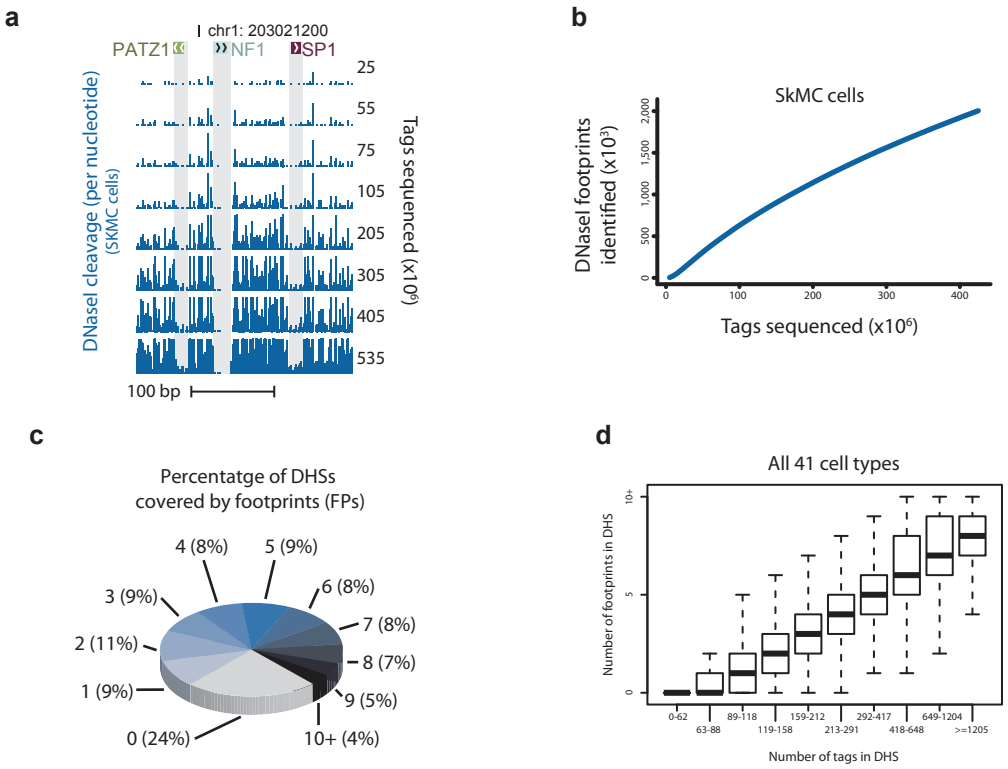
Supplementary Figures

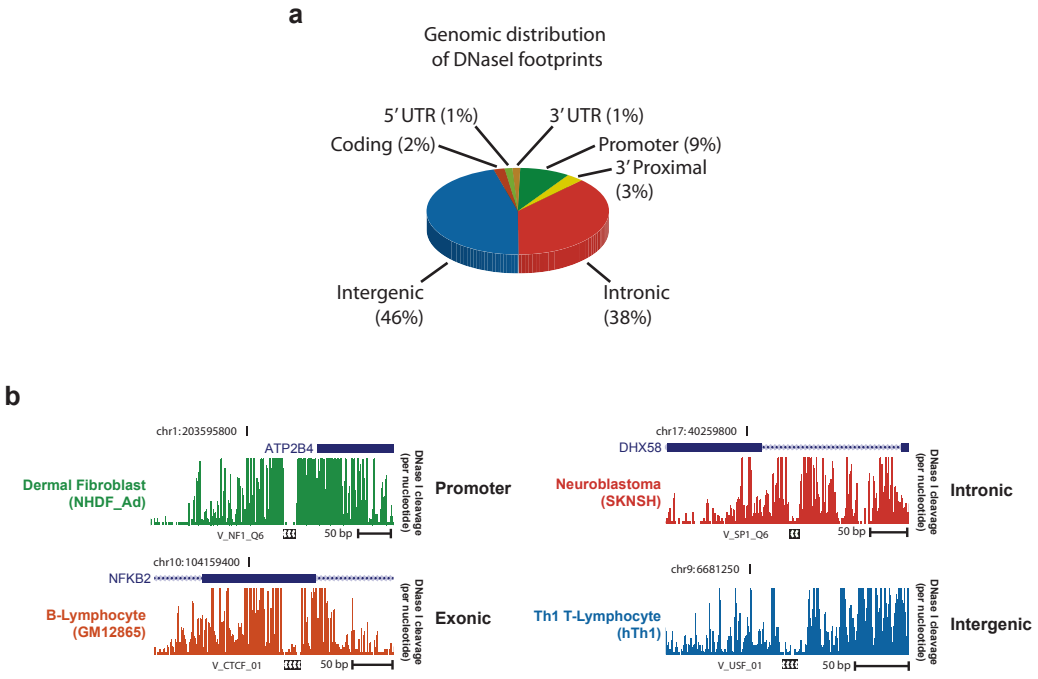
Item	Descriptive title
Supplementary Fig. 1	Identification and distribution of DNaseI footprints
Supplementary Fig. 2	Distribution of DNaseI footprints
Supplementary Fig. 3	Motif density in DNaseI footprints
Supplementary Fig. 4	Association of footprint, occupancy and sequence conservation
Supplementary Fig. 5	Validation of footprints as potential sites of protein occupancy in vitro
Supplementary Fig. 6	DNaseI footprints mark sites of functional *in vivo* protein occupancy
Supplementary Fig. 7	Stereotyped cleavage patterns for different TFs
Supplementary Fig. 8	Anti-correlation of conservation and DNaseI cleavage
Supplementary Fig. 9	General transcriptional activators occupy the PIC footprint
Supplementary Fig. 10	Occupancy of transcription factors differs by mode of interaction with chromatin
Supplementary Fig. 11	Distribution of indirect binding by transcription factor
Supplementary Fig. 12	Distribution of direct and indirect transcription factor binding
Supplementary Fig. 13	Directly bound promoter elements mediate indirect transcription factor interactions
Supplementary Fig. 14	De novo motif discovery in footprints
Supplementary Fig. 15	Conservation and selection of DNaseI footprints

Supplementary Tables

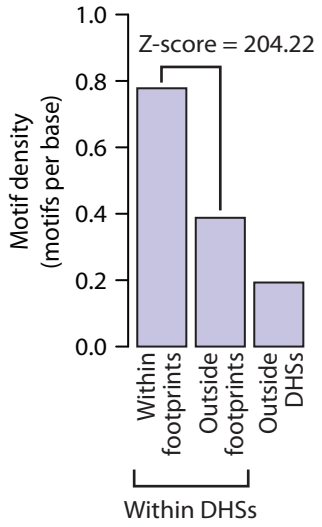
Item	Descriptive title
Supplementary Table 1	Mapping and footprinting statistics for 41 cell lines used in this study
Supplementary Table 2	Summary of footprints within DHSs
Supplementary Table 3	Sequence oligos used for DIPP
Supplementary Table 4	Complete Genomics genome sequence IDs



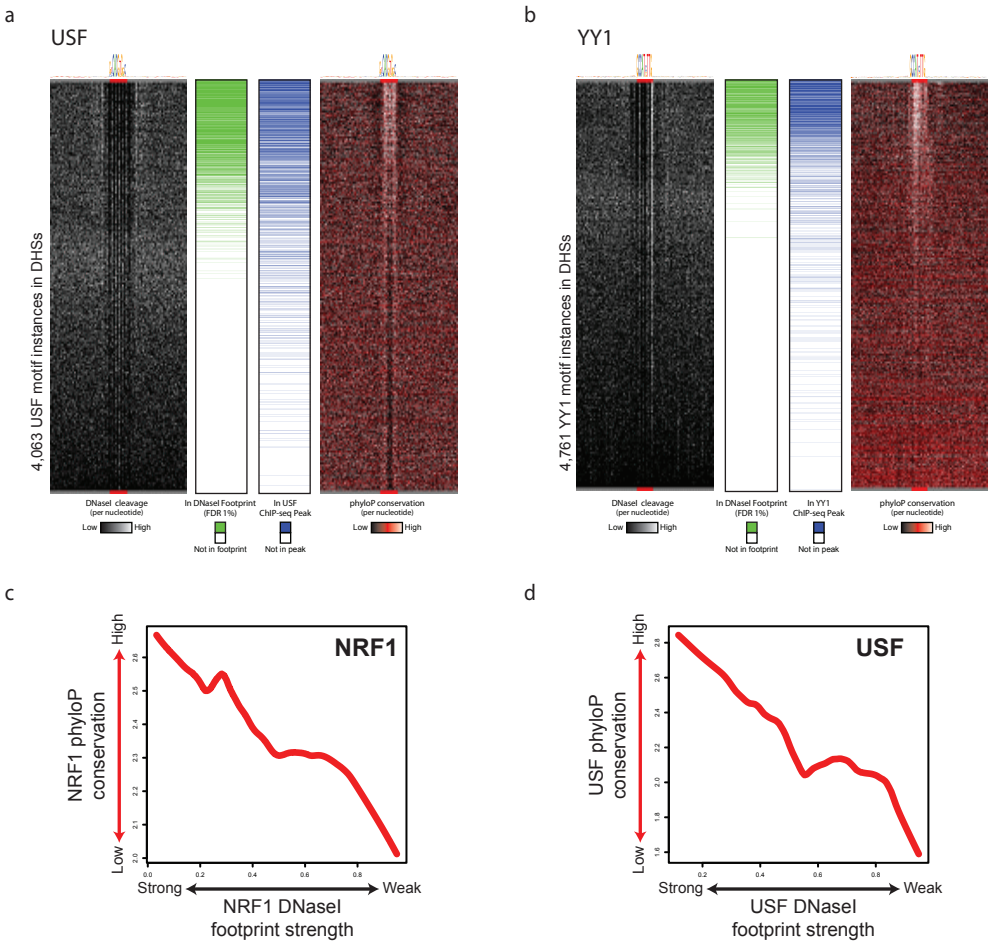
Supplementary Figure 1. Identification and distribution of DNaseI footprints. a, As more DNaseI cleavages are sequenced from SKMC cells, individual DNaseI footprints are easier to distinguish. **b,** The number of DNaseI footprints identified in SKMC cells at varying DNaseI cleavage tag sequencing levels. **c-d,** The number of footprints in DHSs is higher for DHSs with more mapped DNaseI cleavages. DHSs from all 41 cell types were broken into deciles based on the sequencing depth of that DHS. The number of mapped DNaseI cleavages for DHSs in each quantile is indicated below the graph. The box-and-wisker plot shows the distribution of the number of footprints within DHSs for each quantile.



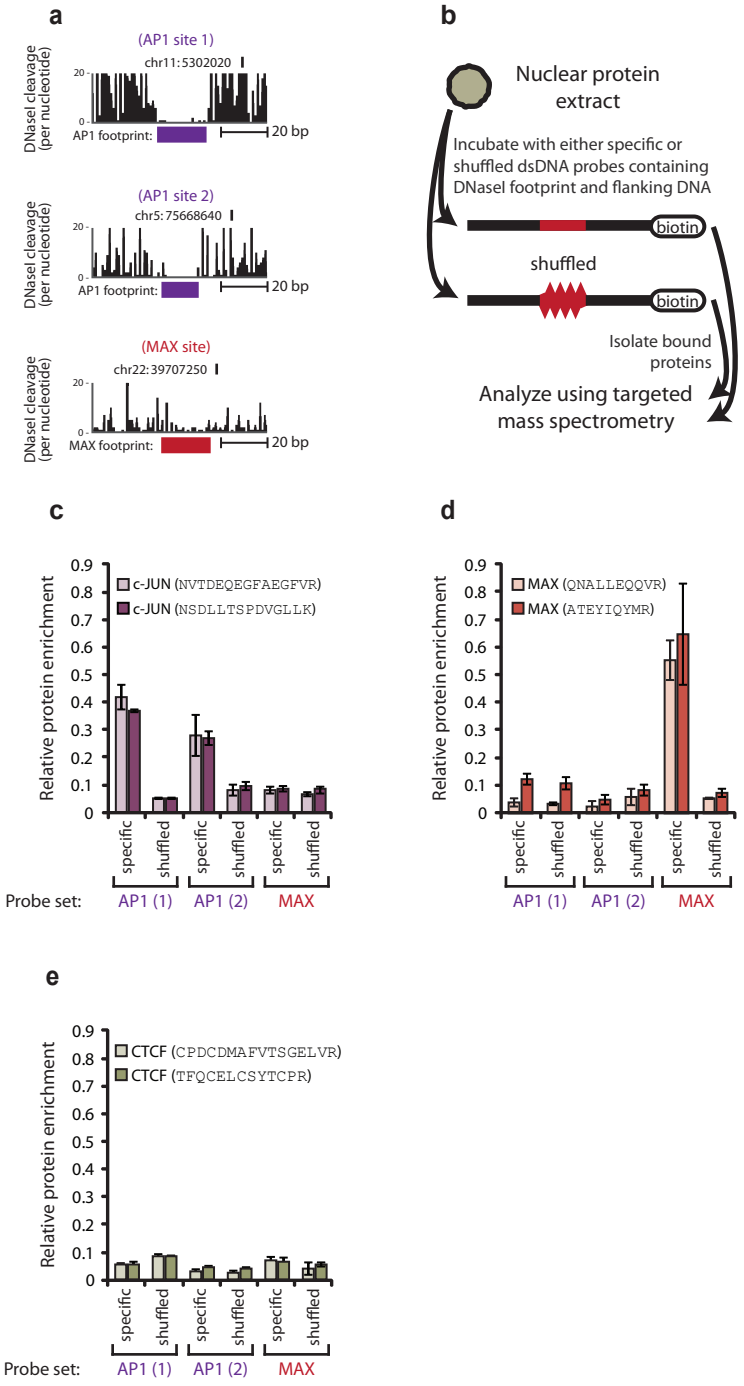
Supplementary Figure 2. Distribution of DNaseI footprints. a, The genomic distribution of footprints found in 41 cell types in relation to annotated genomic features. **b,** Examples of DNaseI footprints at different genomic features.



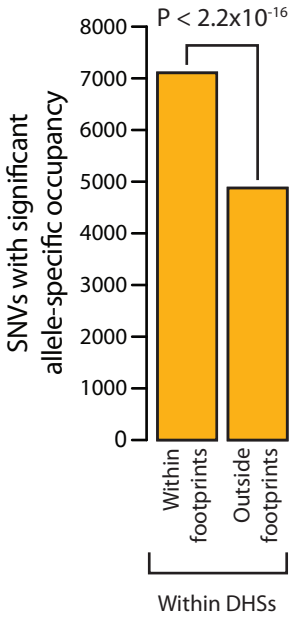
Supplementary Figure 3. Motif density in DNaseI footprints. The density of motifs in DNaseI footprints, DHSs (but not in footprints) and non-hypersensitive genomic regions. Motifs are significantly enriched in footprints (Z-score = 204.22, Genome Structure Correction program comparing the locations of TRANSFAC motifs in 1% FDR footprints).



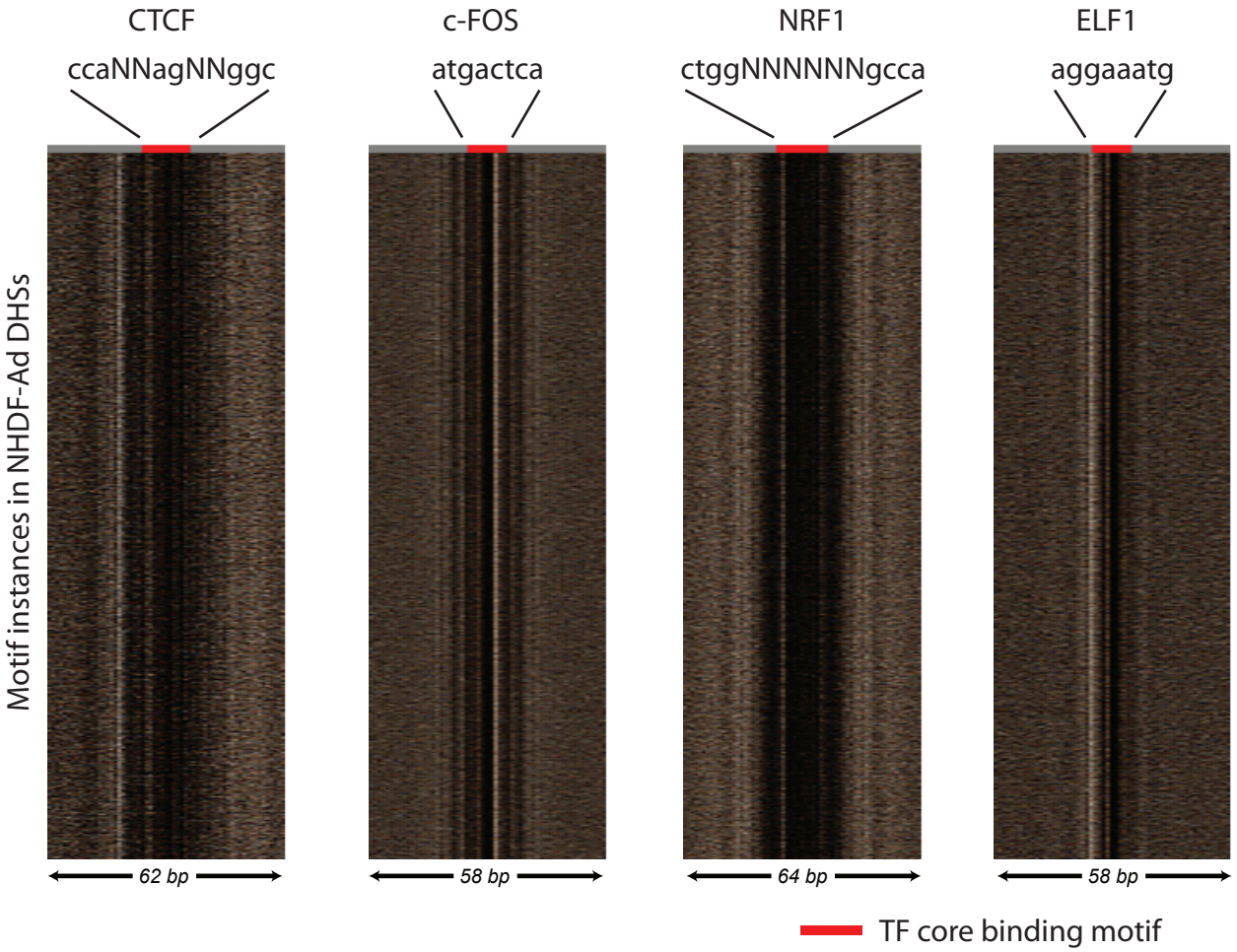
Supplementary Figure 4. Association of footprint, occupancy and sequence conservation. a-b, Heatmaps showing per nucleotide DNaseI cleavage (left) and vertebrate conservation by phyloP (right) for USF (a) and YY1 (b) motifs within K562 DHSs ranked by tag density. Green and blue indicator ticks in the middle indicate the presence of DNaseI footprints and ChIP-seq peaks, respectively, at putative genomic binding sites. **c-d,** Lowess regression of NRF1 (c) and USF (d) maximum phyloP score versus DNaseI footprinting occupancy (footprint occupancy score) at K562 DNaseI footprints marked by NRF1 (c) and USF (d) motifs.



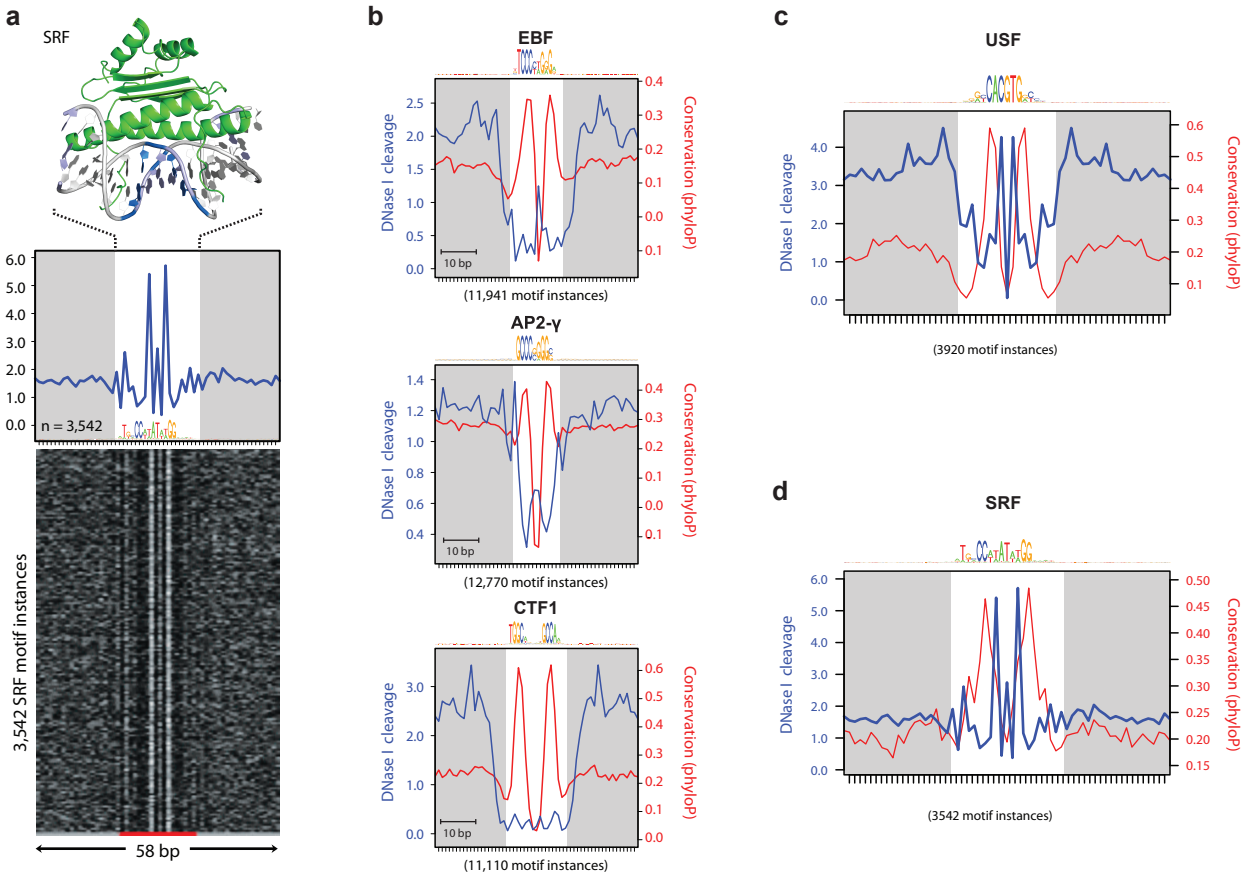
Supplementary Figure 5. Validation of footprints as potential sites of protein occupancy *in vitro*. a, Three genomic loci of varying footprint strength targeted using DNA interacting protein precipitation (DIPP). **b,** Schematic overview of the DIPP protocol. **c-d,** Targeted mass spectrometry measurements of the proteins enriched using the different probe sets. The AP1 protein c-Jun was enriched specifically using the AP1 probes (c) and MAX was enriched specifically using the MAX probe (d). **e,** As a negative control for DIPP, we tested for CTCF binding to the six probes. CTCF did not appear to be enriched in any of the pulldowns.



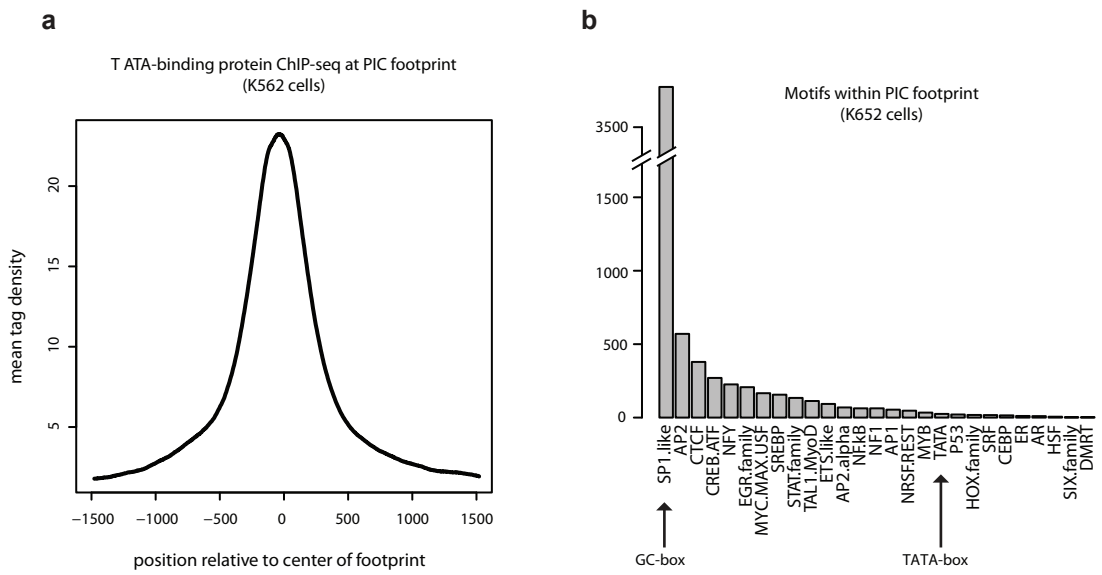
Supplementary Figure 6. DNaseI footprints mark sites of functional *in vivo* protein occupancy. Heterozygous SNVs associated with allele-specific occupancy are significantly enriched inside footprints compared to the rest of the DHS (P < 2.2x10-16, Fisher's exact test).



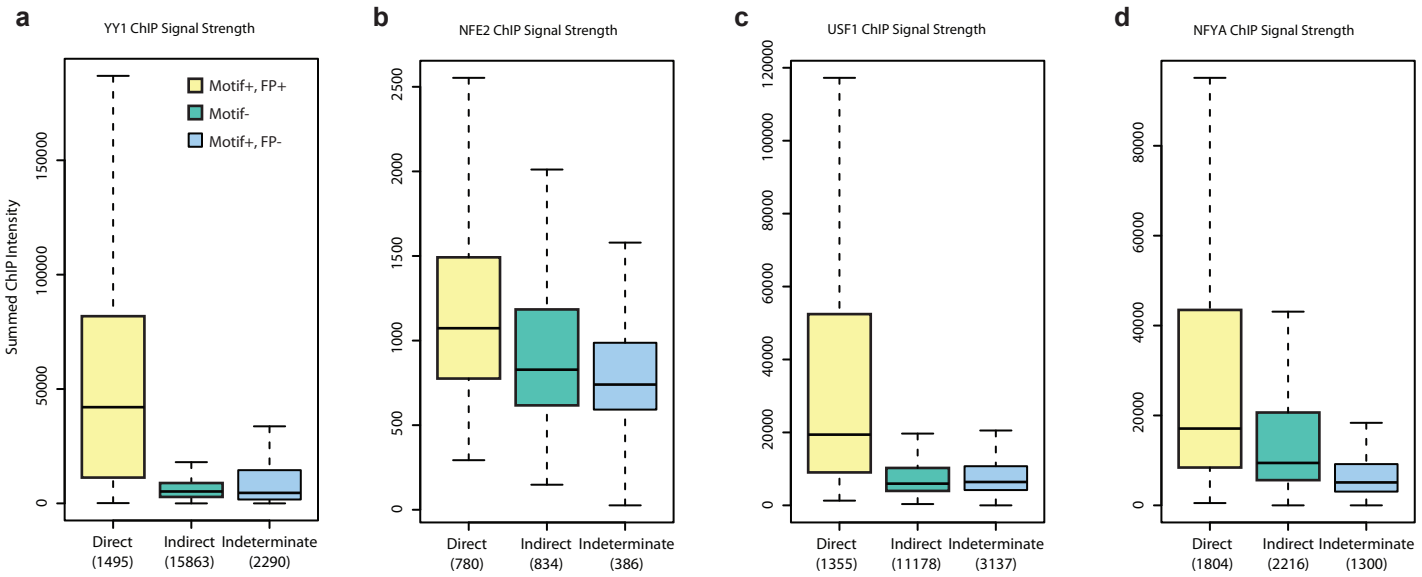
Supplementary Figure 7. Stereotyped cleavage patterns for different TFs. The per-nucleotide DNaseI cleavage patterns at motif instances of 4 different transcription factors in adult dermal fibroblasts (NHDF-Ad). The different motif instances (rows) are randomly ordered.



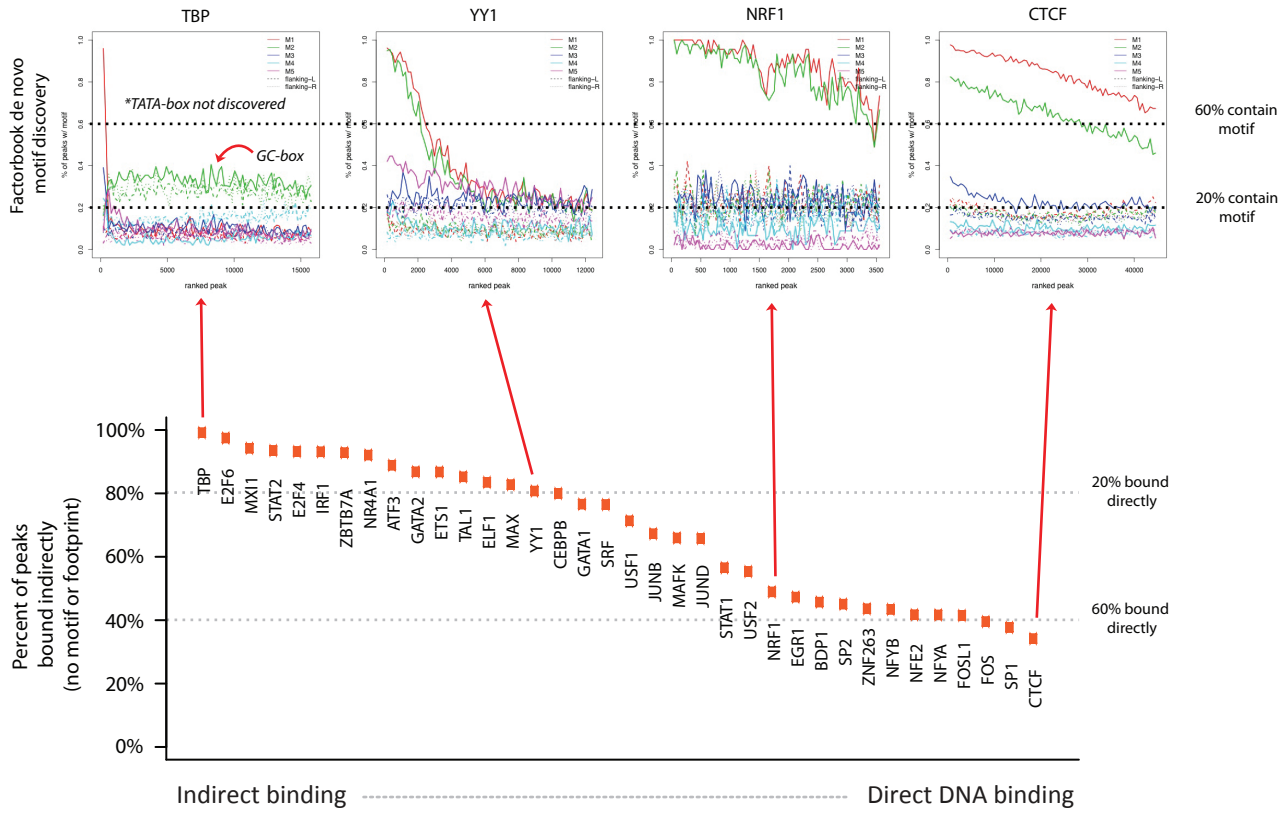
Supplementary Figure 8. Anti-correlation of conservation and DNaseI cleavage for factors with structural data.
a, Similar to Fig. 3a, the co-crystal structure of Serum Response Factor (SRF) bound to its DNA ligand is juxtaposed above the average nucleotide-level DNaseI cleavage pattern (blue) at motif instances of SRF in DNaseI footprints. Nucleotides that are sensitive to cleavage by DNaseI are colored as blue on the co-crystal structure. The motif logo generated from SRF DNaseI footprints is displayed below the DNaseI cleavage pattern. Below is a randomly ordered heatmap showing the per-nucleotide DNaseI cleavage for each motif instance of SRF in DNaseI footprints. **b,** The per-base DNaseI hypersensitivity (blue) and vertebrate phylogenetic conservation (red) for all DNaseI footprints in dermal fibroblasts matching three well annotated transcription factor motifs. The white box indicates width of consensus motif. The number of motif occurrences within DNaseI footprints in indicated below each graph. **c-d,** Cleavage profiles mirror the protein structure and are anti-correlated with vertebrate conservation for USF (c) and SRF (d).



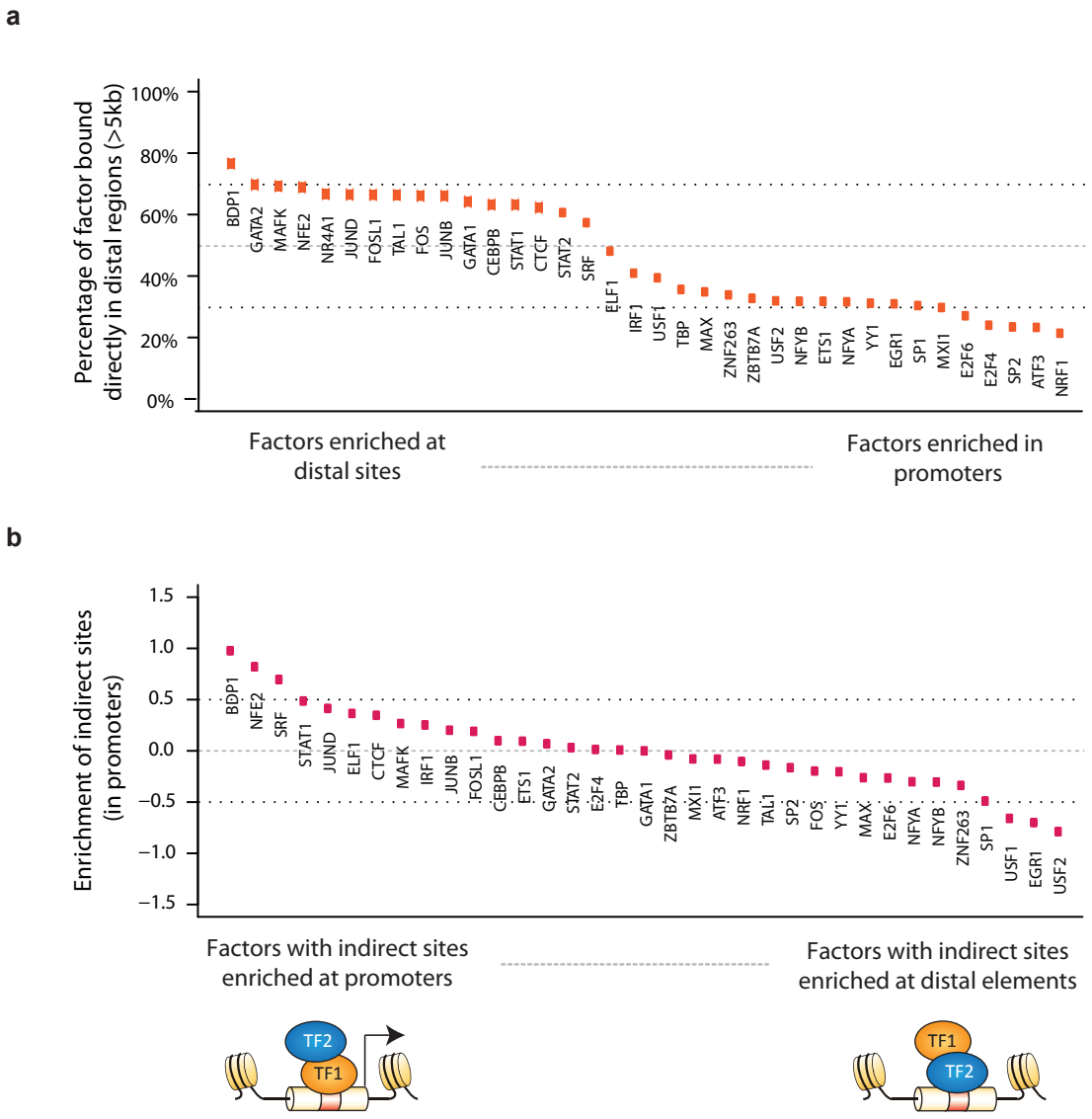
Supplementary Figure 9. General transcriptional activators occupy the PIC footprint. a, Mean ChIP-seq tag density for TATA-binding protein centered on the TSS-linked footprint in K562 cells. **b,** Motifs associated with general transcription factors are found within the footprint. TRANSFAC motifs, reduced by similarity and non-overlapping instances of each motif group, were enumerated inside of the PIC footprint.



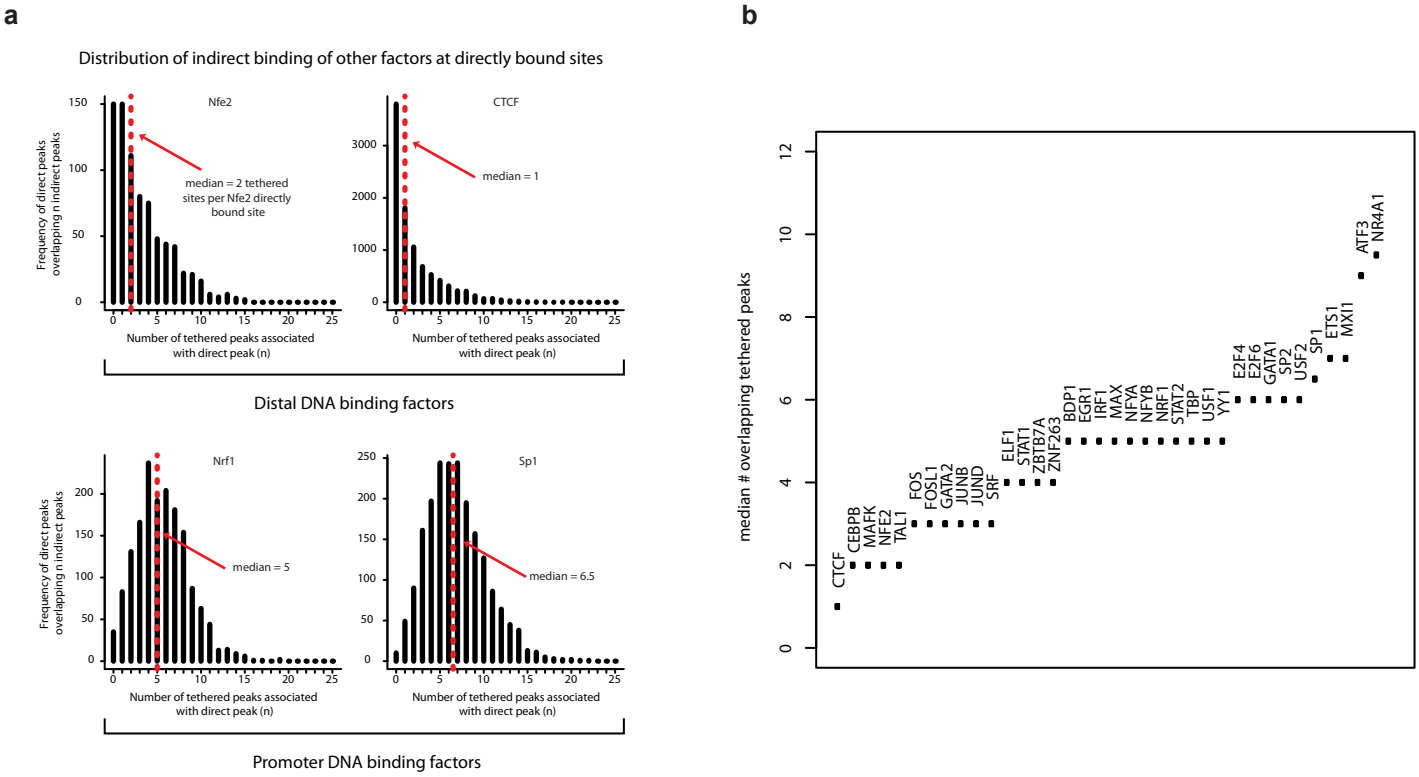
Supplementary Figure 10. Occupancy of transcription factors differs by mode of interaction with chromatin. a-d, ChIP-seq peaks of the factors YY1 (a), NFE2 (b), USF1 (c), and NFYA (d) were partitioned into three classes: direct (footprinted motif; yellow), indirect (no motif; green) and indeterminate (motif with no footprint; blue). The sum of the raw sequencing tags is displayed for each instance of a ChIP-seq peak in each partition. The number of ChIP-seq peaks contributing to each partition is displayed below.



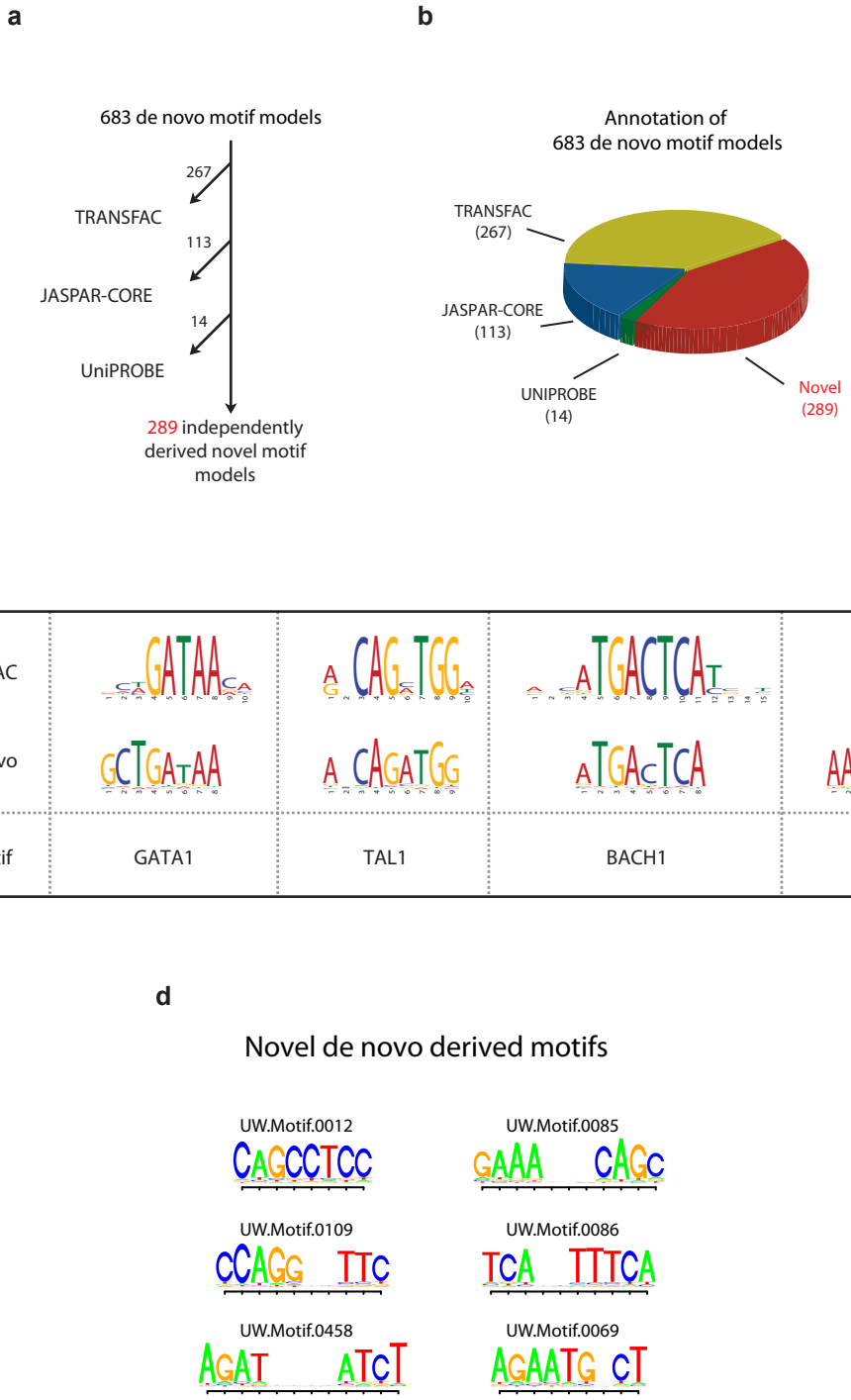
Supplementary Figure 11. Distribution of indirect binding by transcription factor. Transcription factors are ordered by the percentages of total peaks bound indirectly (bottom). We compare our values of indirect binding to motif occurences (presumably direct binding) determined by Factorbook (http://www.factorbook.org) (top). ChIP-seq peaks are ordered by intensity and binned into groups of 500 peaks (x-axis). The fraction of ChIP-seq peaks containing a discovered motif (y-axis) is plotted. Red and green lines represent the known binding motif, except for TATA-binding protein, for which a TATA-box was not identified. The dotted horizontal line on the bottom plot represents 20% and 60% direct binding (80% and 40% indirect, respectively). Corresponding dotted lines are drawn on the Factorbook plots highlighting the percentage of binding sites that contain a cognate recognition site.



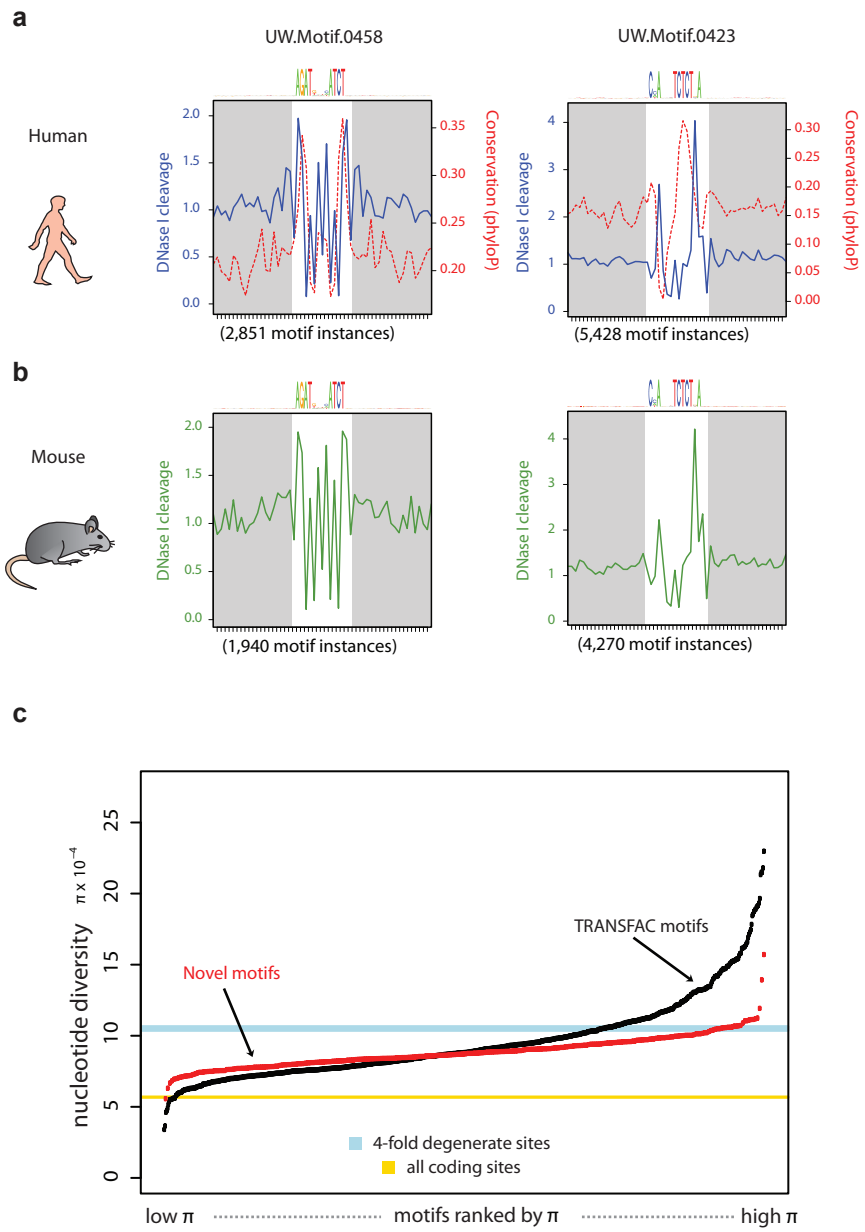
Supplementary Figure 12. Distribution of direct and indirect transcription factor binding. a, The percentage of K562 ChIP-seq peaks bound directly in distal regions was computed for each factor. Here, we define distal as sites greater that 5 kilobases from any GENCODE level 1 and 2 annotated promoter. **b,** The enrichment of indirect ChIP-seq peaks found in promoters for transcription factors in (a). The enrichment is defined as the $\log_2$ ratio between the fraction of indirect sites in promoters and distal regions.



Supplementary Figure 13. Directly bound promoter elements mediate indirect transcription factor interactions. a, The number of overlapping indirect ChIP-seq peaks of other factors was computed for each directly bound ChIP-seq peak and represented as a histogram. On average, directly bound NFE2 ChIP-seq peaks overlap two indirect peaks of other factors, while Sp1 overlaps on average 6.5 indirect peaks. **b,** The median value of overlapping indirect peaks at directly bound sites was computed for many factors.



Supplementary Figure 14. De novo motif discovery in footprints. a, Diagram of the depletion scheme used to identify novel motifs. 683 motifs were filtered in successive order using TOMTOM with TRANSFAC, JASPAR-CORE and UniPROBE. The numbers on the arrows display the number of de novo motifs matched to the corresponding database. **b,** Pie chart annotating the partition of de novo motifs into known and novel motifs. **c,** Example consensus logos of de novo derived motifs that match TRANSFAC models. **d,** Example consensus logos of novel de novo derived motifs using DNaseI footprints.



Supplementary Figure 15. Conservation and selection of DNaseI footprints. a, Phylogenetic conservation (red dashed) and per-base DNaseI hypersensitivity (blue) for all DNaseI footprints in dermal fibroblast cells matching two novel de novo-derived motifs. The white box indicates width of consensus motif. **b,** Per-nucleotide mouse liver DNaseI cleavage patterns at occurrences of the motifs in (a) at DNaseI footprints identified in mouse liver. **c,** The average human nucleotide diversity (π, y-axis) across all motif instances within DNaseI footprints is plotted for each of the motif models in the TRANSFAC database (black, ordered by mean π) and for each of the novel de novo-derived motif models (red, ordered by mean π). Blue bar indicates the average nucleotide diversity (pi) at 4-fold degenerate coding sites (width is equal to 95% confidence interval); gold bar indicates pi at all coding sites (width is equal to 95% confidence interval).

Supplementary Table 1. Mapping and footprinting statistics for 41 cell lines used in this study.

Cell type	Total sequencing reads	Uniquely mapping sequencing tags	SPOT score	Hotspot Regions (Z-score > 2)	Hotspot Region bases (Z-score > 2)	Hotspot Regions (FDR 1%)	Hotspot Region bases (FDR 1%)	Number of Footprints	Footprint bases
AG10803	369,891,377	284,236,136	0.72	251,730	127,453,885	167,583	67,986,704	1,106,404	16,098,123
AoAF	393,665,970	331,043,131	0.68	247,711	157,461,278	149,702	68,131,539	1,566,170	22,727,963
CD20+	300,895,833	240,594,387	0.57	176,008	143,218,339	74,681	55,800,743	603,190	8,277,635
CD34+ Mobilized	287,605,852	221,098,234	0.7	206,033	119,621,324	125,354	59,782,641	902,386	13,494,265
fBrain	247,674,296	202,264,605	0.72	282,491	145,389,927	166,439	72,393,628	1,022,782	16,300,790
fHeart	328,340,544	264,719,957	0.55	292,700	200,606,802	149,848	77,501,676	954,914	14,276,245
fLung	318,544,188	268,295,068	0.67	338,476	206,554,970	184,428	89,331,237	1,181,235	18,884,673
GM06990	238,189,351	137,532,640	0.62	194,407	124,993,150	71,679	45,274,471	434,561	6,588,543
GM12865	375,878,018	263,505,515	0.56	227,299	132,981,145	118,973	60,969,626	811,374	11,983,394
HAEpiC	347,415,753	281,238,046	0.76	294,231	153,728,366	189,376	78,262,576	1,506,475	20,870,168
HA-h	463,552,126	367,030,503	0.53	316,350	215,720,624	183,402	96,933,344	966,188	13,539,911
HCF	375,266,490	289,961,235	0.7	251,268	136,811,533	158,618	69,285,351	1,057,743	15,531,256
HCM	348,691,463	274,790,101	0.73	259,190	148,481,494	176,297	80,809,789	1,130,292	16,902,956
HCPEpiC	305,511,987	253,925,492	0.76	304,490	166,336,755	194,361	80,354,004	1,296,454	18,684,469
HEEpiC	472,469,677	342,975,637	0.58	326,246	170,658,884	195,205	76,649,605	1,263,648	18,866,093
HepG2	248,320,583	168,883,956	0.57	199,174	115,287,739	73,091	35,775,390	448,678	6,938,557
H7-hESC	401,363,495	302,050,785	0.61	491,178	236,940,779	248,021	99,574,058	1,279,454	18,940,427
HFF	344,017,295	262,521,646	0.59	271,655	158,223,732	173,197	85,614,520	590,904	8,338,681
HIPEpiC	333,832,037	254,744,863	0.58	331,341	179,188,377	209,537	82,901,328	1,089,936	16,510,277
HMF	384,696,734	311,000,443	0.75	266,757	137,564,020	175,343	71,776,020	1,434,330	19,950,879
HMVEC-dBl-Ad	292,823,995	239,063,258	0.72	194,045	124,284,025	142,138	75,710,345	1,085,741	14,936,244
HMVEC-dBl-Neo	368,114,784	293,473,622	0.57	198,849	142,843,889	145,392	87,577,362	1,061,860	15,410,057
HMVEC-dLy-Neo	338,875,033	270,345,138	0.58	196,251	122,494,916	132,021	67,831,113	989,626	14,367,015
HMVEC-LLy	417,841,726	313,021,953	0.62	176,634	115,933,614	122,591	65,162,445	872,721	12,565,891
HPAF	320,990,135	255,470,482	0.7	256,698	140,840,570	169,984	81,414,365	1,090,215	15,983,810
HPdLF	371,416,176	304,268,872	0.67	266,670	168,672,350	156,380	73,248,071	1,404,872	20,203,519
HPF	368,365,528	296,713,698	0.66	235,885	138,436,787	138,004	63,605,792	1,175,289	17,027,101
HRCEpiC	284,056,343	236,736,388	0.6	307,274	155,486,022	178,791	71,906,908	1,187,325	17,124,566
HSMM	467,134,471	367,269,086	0.66	331,104	177,231,683	215,419	94,115,944	1,668,243	23,986,641
Th1	232,708,777	171,609,858	0.64	154,717	117,670,939	63,672	43,580,258	498,505	7,448,335
HVMF	341,994,992	279,802,866	0.63	265,941	159,102,198	154,706	72,203,801	1,263,833	18,402,231
IMR90	309,171,904	242,507,116	0.53	286,260	141,276,695	184,888	79,251,166	970,277	14,355,207
K562	268,452,588	179,970,820	0.56	256,735	157,203,075	125,859	64,943,646	498,683	7,161,934
NB4	404,801,445	323,812,091	0.56	236,509	141,522,055	119,640	62,875,330	1,049,300	15,418,984
NH-A	307,812,903	231,589,045	0.57	280,019	148,952,898	176,271	80,278,529	977,923	14,329,589
NHDF-Ad	300,516,213	235,650,107	0.81	296,898	151,532,500	212,841	93,354,642	1,429,399	20,950,088
NHDF-neo	482,603,639	373,361,757	0.7	275,166	153,433,828	172,878	76,999,986	1,532,853	22,147,781
NHLF	454,391,713	357,163,548	0.71	294,352	166,314,415	190,888	85,453,334	1,567,106	22,751,625
SAEC	296,719,796	243,838,476	0.58	291,390	159,165,382	184,542	72,503,786	1,256,188	18,742,067
SKMC	632,856,867	543,886,965	0.81	311,537	158,366,070	193,202	74,105,557	2,370,723	31,607,291
SK-N-SH_RA	217,691,024	164,615,431	0.7	160,880	91,155,430	70,493	37,635,241	498,926	7,609,202

Supplementary Table 2. Summary of footprints within DHSs.

Cell type	Total FPs	Total DHS peaks	FPs in DHS peaks	DHS peaks with FP	Mean FP per DHS peak
AG10803	1,106,404	181,473	677,479	139,806	4.85
AoAF	1,566,170	165,258	820,187	148,612	5.52
CD20+	603,190	104,139	303,432	72,752	4.17
CD34+ Mobilized	902,386	147,098	560,210	117,862	4.75
fBrain	1,022,782	182,501	636,950	140,256	4.54
fHeart	954,914	173,135	562,780	129,032	4.36
fLung	1,181,235	205,880	681,428	160,948	4.23
GM06990	434,561	92,709	195,168	49,295	3.96
GM12865	811,374	143,716	487,801	104,614	4.66
HAEpiC	1,506,475	205,033	913,983	172,375	5.3
HA-h	966,188	200,014	506,977	134,600	3.77
HCF	1,057,743	174,667	647,025	135,144	4.79
HCM	1,130,292	193,375	696,405	146,587	4.75
HCPEpiC	1,296,454	210,380	826,565	167,674	4.93
HEEpiC	1,263,648	209,838	834,743	173,806	4.8
HepG2	448,678	90,775	228,280	54,600	4.18
H7-hESC	1,279,454	266,618	808,678	189,181	4.27
HFF	590,904	192,282	384,995	106,555	3.61
HIPEpiC	1,089,936	225,744	731,881	164,569	4.45
HMF	1,434,330	190,512	874,301	162,132	5.39
HMVEC-dBl-Ad	1,085,741	162,593	644,136	123,503	5.22
HMVEC-dBl-Neo	1,061,860	168,436	633,452	124,918	5.07
HMVEC-dLy-Neo	989,626	153,107	603,547	120,801	5
HMVEC-LLy	872,721	144,886	550,573	111,126	4.95
HPAF	1,090,215	188,071	684,069	140,068	4.88
HPdLF	1,404,872	171,349	785,700	147,294	5.33
HPF	1,175,289	154,397	683,890	131,805	5.19
HRCEpiC	1,187,325	192,147	723,271	146,937	4.92
HSMM	1,668,243	228,282	937,370	184,856	5.07
Th1	498,505	84,201	220,748	53,494	4.13
HVMF	1,263,833	170,340	688,248	137,947	4.99
IMR90	970,277	199,752	646,563	139,353	4.64
K562	498,683	142,986	305,128	72,048	4.24
NB4	1,049,300	143,838	588,282	117,445	5.01
NH-A	977,923	191,510	601,546	130,914	4.59
NHDF-Ad	1,429,399	230,696	891,028	179,529	4.96
NHDF-neo	1,532,853	187,962	840,887	160,662	5.23
NHLF	1,567,106	206,254	896,218	173,139	5.18
SAEC	1,256,188	198,442	791,686	160,216	4.94
SKMC	2,370,723	205,493	1,230,494	198,952	6.18
SK-N-SH_RA	498,926	89,968	259,755	61,111	4.25

Supplementary Table 3. Sequence oligos used for DIPP.

Probe ID	IDT Sequence
MAX Specific F	/5BiosG/CTGGAGACTTGCAGGGTGGACAGACACACGTGGGGAAGGTTCCCGCTGCACACAACTCAACTCTGACCTG
MAX Specific R	CAGGTCAGAGTTGAGTTGTGTGCAGCGGGAACCTTCCCCACGTGTGTCTGTCCACCCTGCAAGTCTCCAG
MAX Shuffled F	/5BiosG/CTGGAGACTTGCAGGGTGGACAGACACAGTTCAGAGCGGCCGTGGCTGCACACAACTCAACTCTGACCTG
MAX Shuffled R	CAGGTCAGAGTTGAGTTGTGTGCAGCCACGGCCGCTCTGAACTGTGTCTGTCCACCCTGCAAGTCTCCAG
AP1 Site 1 Specific F	/5BiosG/CATCTGGGCACACACCCTAAGCCTCAGCATGACTCATCATGACTCAGCATTGCTGTGCTTGAGCCAGAAG
AP1 Site 1 Specific R	CTTCTGGCTCAAGCACAGCAATGCTGAGTCATGATGAGTCATGCTGAGGCTTAGGGTGTGTGCCCAGATG
AP1 Site 1 Shuffled F	/5BiosG/CATCTGGGCACACACCCTAAGCCTTGCAACCGGACAACAAGCGCCTTATTTTCTGTGCTTGAGCCAGAAG
AP1 Site 1 Shuffled R	CTTCTGGCTCAAGCACAGAAAATAAGGCGCTTGTTGTCCGGTTGCAAGGCTTAGGGTGTGTGCCCAGATG
AP1 Site 2 Specific F	/5BiosG/TGGGATTATCAGGCTGGAGTTCTCTGTCATTAGGATGACTCATCATTTTTCTATCTCTGCTTCCATTGCT
AP1 Site 2 Specific R	AGCAATGGAAGCAGAGATAGAAAAATGATGAGTCATCCTAATGACAGAGAACTCCAGCCTGATAATCCCA
AP1 Site 2 Shuffled F	/5BiosG/TGGGATTATCAGGCTGGAGTTCTCTGTAGTATATTCCTCATATTCTTGAGCTATCTCTGCTTCCATTGCT
AP1 Site 2 Shuffled R	AGCAATGGAAGCAGAGATAGCTCAAGAATATGAGGAATATACTACAGAGAACTCCAGCCTGATAATCCCA

Supplementary Table 4. Complete Genomics genome sequence IDs.

Population	Assemblies
African	GS19238-1100-37-ASM
	GS19239-1100-37-ASM
	GS19017-1100-37-ASM
	GS19020-1100-37-ASM
	GS19025-1100-37-ASM
	GS19026-1100-37-ASM
	GS21732-1100-37-ASM
	GS21733-1100-37-ASM
	GS21767-1100-37-ASM
	GS18501-1100-37-ASM
	GS18502-1100-37-ASM
	GS18504-1100-37-ASM
	GS18505-1100-37-ASM
	GS18508-1100-37-ASM
	GS18517-1100-37-ASM
	GS19219-1100-37-ASM
African - American	GS19700-1100-37-ASM
	GS19701-1100-37-ASM
	GS19703-1100-37-ASM
	GS19704-1100-37-ASM
	GS19834-1100-37-ASM
Asian	GS18940-1100-37-ASM
	GS18942-1100-37-ASM
	GS18947-1100-37-ASM
	GS18956-1100-37-ASM
	GS18526-1100-37-ASM
	GS18537-1100-37-ASM
	GS18555-1100-37-ASM
	GS18558-1100-37-ASM
European	GS12889-1100-37-ASM
	GS12890-1100-37-ASM
	GS12891-1100-37-ASM
	GS12892-1100-37-ASM
	GS20502-1100-37-ASM
	GS20509-1100-37-ASM
	GS20510-1100-37-ASM
	GS20511-1100-37-ASM
	GS06985-1100-37-ASM
	GS06994-1100-37-ASM
	GS07357-1100-37-ASM
	GS10851-1100-37-ASM
	GS12004-1100-37-ASM
Gujarati	GS20845-1100-37-ASM
	GS20846-1100-37-ASM
	GS20847-1100-37-ASM
	GS20850-1100-37-ASM
Hispanic	HG00731-1100-37-ASM
	HG00732-1100-37-ASM
	GS19735-1100-37-ASM
	GS19648-1100-37-ASM
	GS19649-1100-37-ASM
	GS19669-1100-37-ASM
	GS19670-1100-37-ASM