Supplementary Methods

Comparing motifs

Motif similarity is defined as the maximal Pearson's correlation of the PWM made into a 4xN vector across all offsets and both orientations, padding unmatched positions with N's (102). Motifs are considered a match if they have a similarity of at least 0.75; weak matches or variants are determined through manual inspection.

Collecting known motifs

Known motifs were collected from large scale data sets or databases. We collected human, mouse, and rat motifs from Transfac (version 11.3) (11), vertebrate motifs from Jaspar (version 2008) (12), and large scale systematic motifs generated using Protein Binding Microarrays (13, 14, 15) and high-throughput SELEX (16). HGNC gene names (103) were used to describe the names of motifs whenever possible.

Processing and naming of experimental data sets

Human protein binding peaks (excluding Pol2/Pol3) were taken from the January, 2011 freeze of ENCODE (8). The processing for these data sets is described in (104, 10) and the original data sets are available on the website accompanying this paper. To avoid potentially confounding issues, we excluded from our analysis: (1) the Y and mitochondrial chromosomes; (2) the hg19 rmsk and simple repeat tracks from UCSC (created on April 27, 2009) (105); and (3) protein coding regions, exons for non-coding genes, and 3' untranslated regions taken from GENCODE v4 (106). Like with the motif names, HGNC gene names (103) are used to describe the data sets.

Performing de novo motif discovery

Peaks were randomly partitioned into two data sets for the purpose of separating discovery and enrichment and limit over-fitting. The top 250 peaks for the first partition were used in motif discovery because high intensity peaks often have better enrichment for motifs (101). Five tools were run independently run on each data set: AlignACE (v4.0 with default parameters) (17), MDscan (v2004 with default parameters) (18), MEME (v3.5.7 with a maximum of 10 iterations and -maxw 10, 15, and 25 for the three motifs) (19), Weeder (v1.4.2 with option large) (20), and Trawler (v1.2 with 250 random intergenic blocks for background) (21). Any motifs beyond the top three for any method on one data set were discarded.

Computing enrichments of motifs

Motifs were matched against the genome using a PWM threshold corresponding to a p-value of 4^{-8} as determined by TFM-Pvalue (107), with the intent to imitate the frequency a fully specified 8-mer would match the genome.

Enrichments are always computed by taking a foreground region (i.e. bound regions for a data set) and comparing it to a background region (e.g. Intergenic non-repetitive regions) where regions in the foreground but not in the background are discarded. For a given motif, the fraction of matches that fall within the foreground is computed and divided by the corresponding fraction for shuffled control motifs (control motif generation details in (97)). To prevent spuriously high enrichments due to small counts, we compute a binomial confidence interval (with z=1.5) (108) around both motif and control fractions and take the extreme which leads to the enrichment closest to 1 (the software available on the website implements this enrichment metric).

For each factor group we order the motifs from all discovery programs by the enrichment in the second random partition for the discovery data set. For this step only, we restrict our analysis to 10% of the background regions to reduce the amount of computation. We then select discovered motifs for each factor by this rank order discarding any motif that matches a previous one with similarity greater than 0.75. We supplement these discovered motifs with the known motifs from the literature (described above) and rematch the motifs to the complete background regions and produce comparable enrichments for all data sets.

Supplementary Results

Robustness of motif similarity

While it has been shown that short motifs will sometimes have similar correlations just by chance (109, 110), this happens infrequently for the 0.75 similarity threshold with our database of motifs. When two motifs in the database match, scrambling one of the motifs 100 times leads to more than 5 matches only 6.4% of the time.

Moreover, a given motif in the database is 34.7 times more likely to match another motif in a database than a scramble. While this statistic is hard to interpret due to, on one hand, the significant redundancy in the database and, on the other, the partial exclusion of similar discovered motifs, it demonstrates that spurious motif matches should be relatively uncommon.

Effect of peak intensity on motif enrichment

While it has been shown that motif enrichment varies by peak intensity (101), we find that our results are largely robust to restricting to only the 1,000 strongest peaks per data set (89% of data sets have at least 1,000 peaks). The Pearson correlation of motif enrichment (for all motifs) for one data set when using all peaks or only the top 1,000 has a median of 0.89 (middle 50%: 0.85 to 0.95).

Moreover, the enrichment of the first discovered motif for a factor group has strong correlation between the complete and restrict data sets: for the 55 factor groups with at least two data sets and a discovered motif, the median correlation of the first discovered motif's enrichment is 0.97 (middle 50%: 0.81 to 1.00).

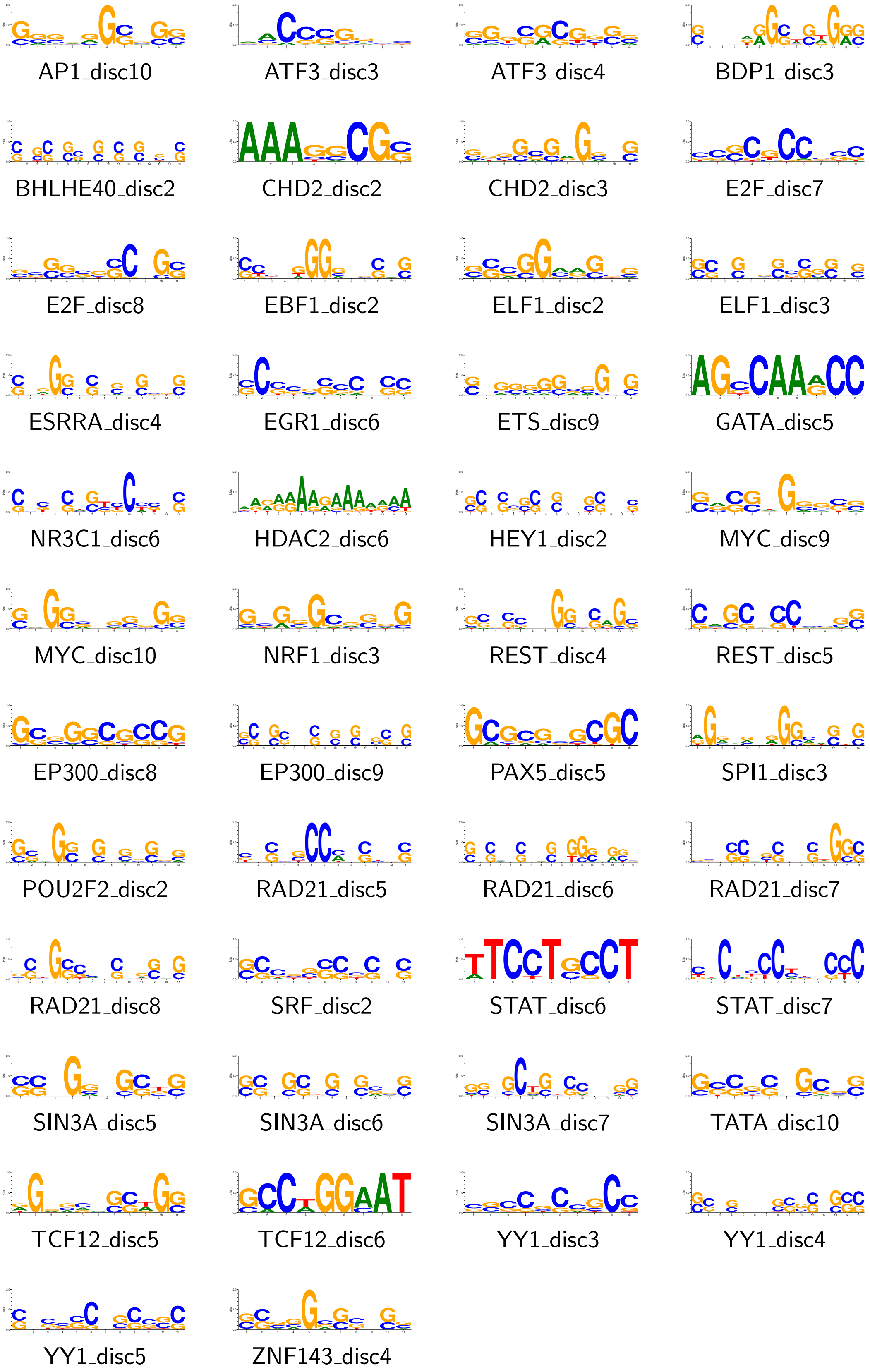


Table S1: Low complexity or weakly enriched motifs.

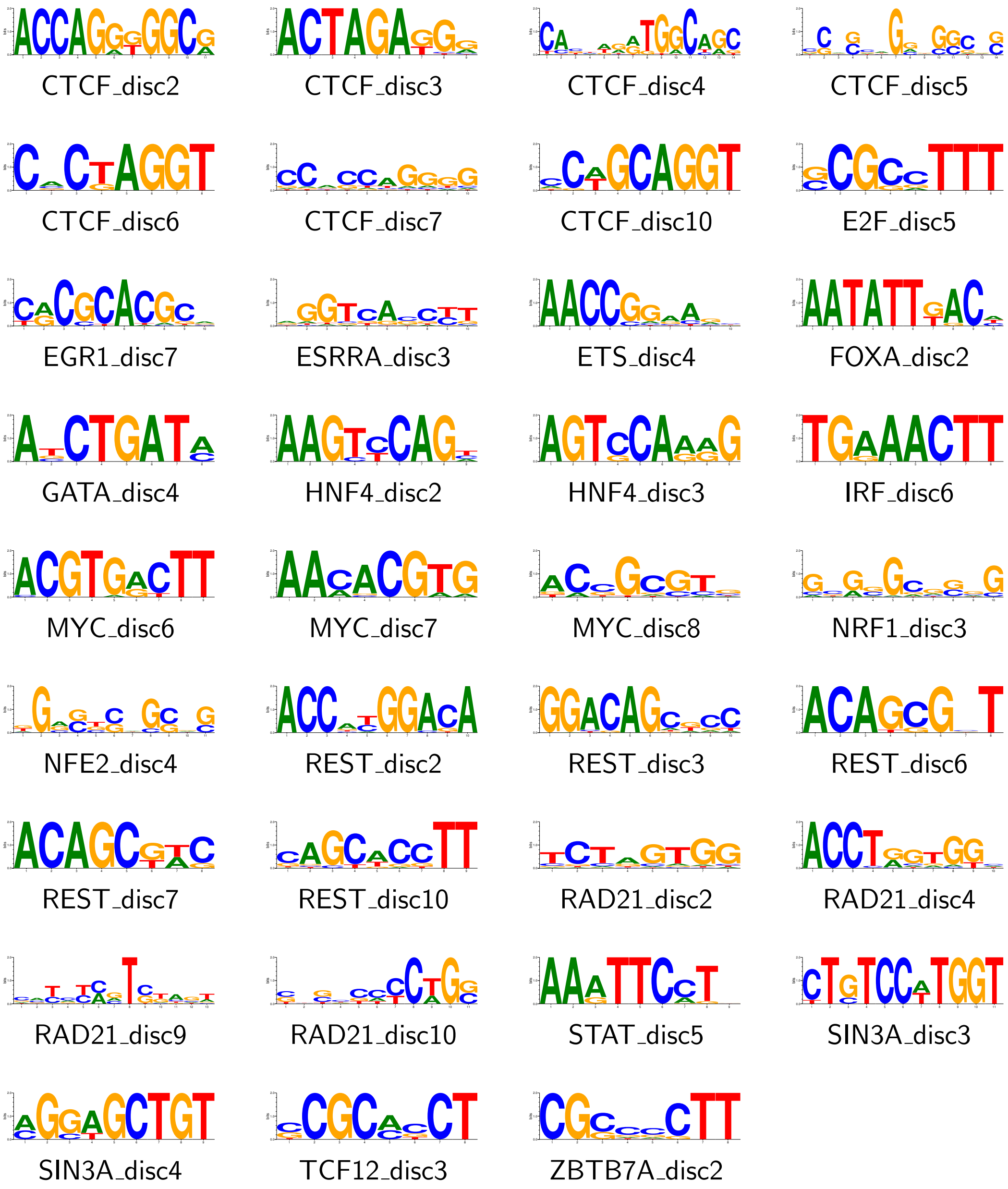
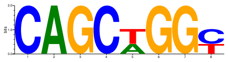
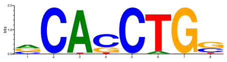
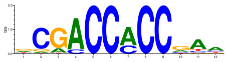
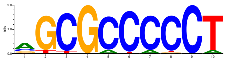
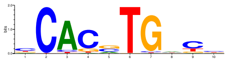
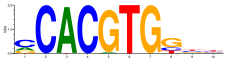
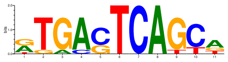
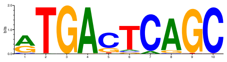
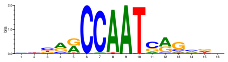
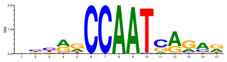
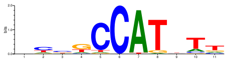
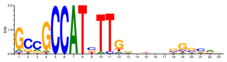
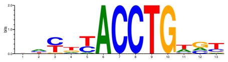
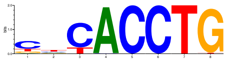
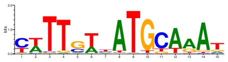
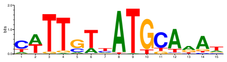


Table S2: Motifs with weak similarity to some other discovered motif. These occur frequently for factors with long known motifs that are amenable to breaking apart. We do not believe that these represent distinct motifs and are likely artifacts of the discovery process.

Known motif	Discovered motif
 TCF12_known1 3.0 (9.5)	 TCF12_disc1* 5.6 (11.2)
 ZBTB7A_known4 1.3 (10.0)	 ZBTB7A_disc1 2.2 (43.5)
 MXI1_known1 3.2 (9.6)	 MXI1_disc2* 4.4 (29.5)
 NFE2_known1 56.4 (78.8)	 NFE2_disc1* 75.9 (125.7)
 NFY_known6 72.6 (110.6)	 NFY_disc1 89.9 (84.7)
 YY1_known5 39.5 (38.6)	 YY1_disc1 47.5 (94.3)
 ZEB1_known1 4.9 (3.4)	 ZEB1_disc1 5.8 (11.1)
 POU5F1_known3 28.3 (22.7)	 POU5F1_disc1* 30.0 (24.8)

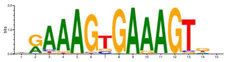
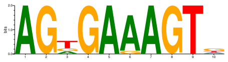
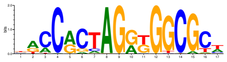
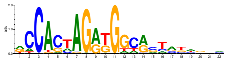
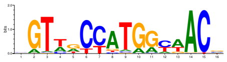
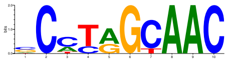
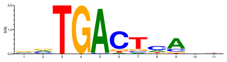
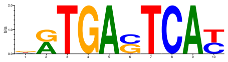
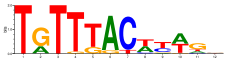
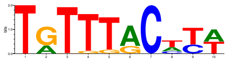
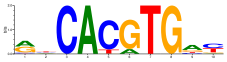
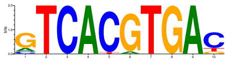
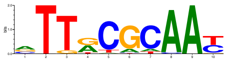
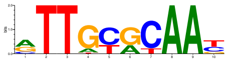
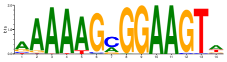
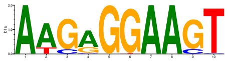
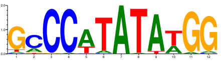
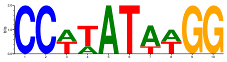
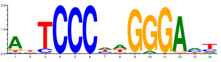
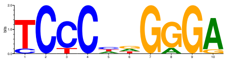
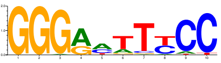
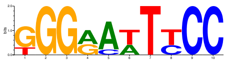
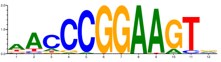
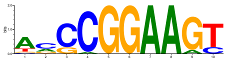
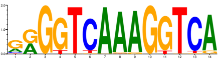
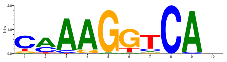
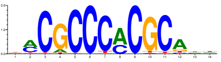
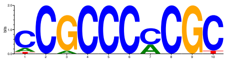
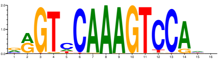
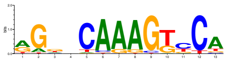
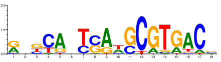
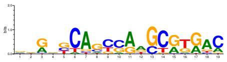
Known motif	Discovered motif
 PRDM1_known2 18.4 (23.3)	 PRDM1_disc1* 19.2 (25.1)
 CTCF_known2 78.0 (219.6)	 CTCF_disc1* 77.3 (118.5)
 RFX5_known6 28.6 (33.8)	 RFX5_disc1* 28.3 (60.7)
 AP1_known3 45.7 (52.5)	 AP1_disc3* 45.1 (82.0)
 FOXA_known6 19.9 (14.4)	 FOXA_disc1* 19.5 (11.2)
 MYC_known22 104.8 (334.2)	 MYC_disc1* 97.8 (139.5)
 CEBPB_known8 84.9 (100.6)	 CEBPB_disc1* 79.0 (91.4)
 SPI1_known4 26.5 (30.5)	 SPI1_disc1* 24.6 (37.0)

Table S3: Complete version of Figure 4, with all factors for which a discovered motif matching a known motif was found. When the discovered motif is not disc1, a better ranking motif was found that did not match a literature motif. * indicates that other discovered motifs were found, even after manual exclusion of weakly redundant and low complexity motifs.

Known motif	Discovered motif
 SRF_known7 126.6 (88.9)	 SRF_disc1 112.1 (93.8)
 EBF1_known4 8.7 (28.0)	 EBF1_disc1 7.3 (22.3)
 NFKB_known3 39.8 (78.5)	 NFKB_disc1 30.3 (45.7)
 ELF1_known2 22.9 (68.8)	 ELF1_disc1 17.0 (105.9)
 RXRA_known10 9.6 (12.4)	 RXRA_disc1* 7.0 (8.2)
 EGR1_known8 11.9 (299.6)	 EGR1_disc1* 8.6 (238.1)
 HNF4_known18 28.0 (34.2)	 HNF4_disc1* 20.0 (30.2)
 PAX5_known5 20.5 (78.2)	 PAX5_disc1* 14.3 (47.3)

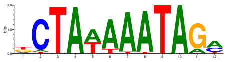
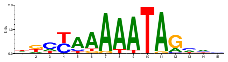
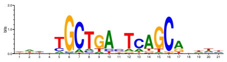
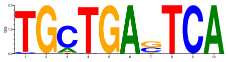
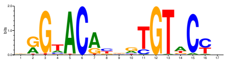
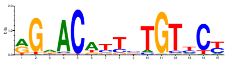
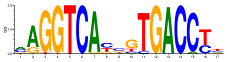
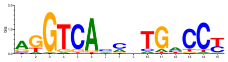
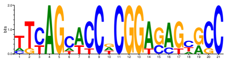
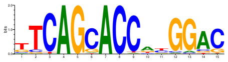
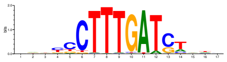
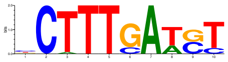
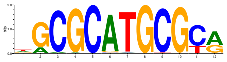
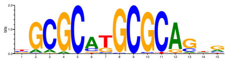
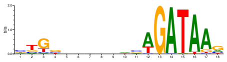
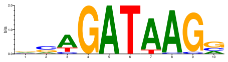
Known motif	Discovered motif
 MEF2_known12 12.0 (6.8)	 MEF2_disc1* 7.9 (6.7)
 MAF_known7 57.8 (68.2)	 MAF_disc1 37.1 (50.1)
 NR3C1_known17 33.5 (49.4)	 NR3C1_disc1* 20.8 (25.0)
 ESRRA_known6 35.6 (60.4)	 ESRRA_disc1* 21.0 (39.0)
 REST_known2 66.9 (188.9)	 REST_disc1 39.4 (115.6)
 TCF7L2_known5 10.9 (7.8)	 TCF7L2_disc2* 6.0 (9.0)
 NRF1_known2 84.7 (1196.5)	 NRF1_disc1* 41.3 (969.7)
 GATA_known14 23.8 (31.3)	 GATA_disc1* 6.9 (23.4)

Table S3 (continued)

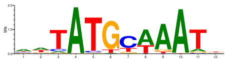
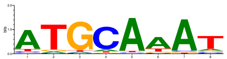
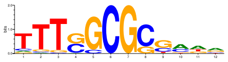
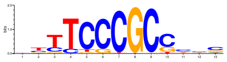
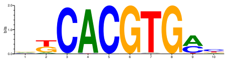
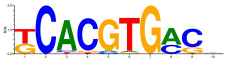
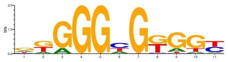
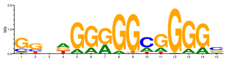
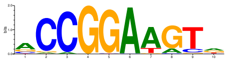
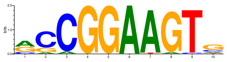
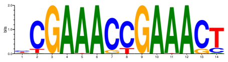
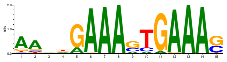
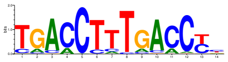
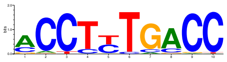
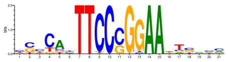
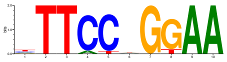
Known motif	Discovered motif
 POU2F2_known15 21.1 (5.3)	 POU2F2_disc1 6.1 (11.0)
 E2F_known9 12.1 (127.5)	 E2F_disc3* 3.1 (71.2)
 BHLHE40_known4 70.2 (168.4)	 BHLHE40_disc1 17.3 (167.0)
 SP1_known8 5.8 (37.7)	 SP1_disc3* 1.0 (8.8)
 ETS_known18 53.9 (306.1)	 ETS_disc2* 8.0 (330.6)
 IRF_known20 61.0 (74.1)	 IRF_disc3* 8.8 (15.4)
 NR2C2_known1 53.4 (51.8)	 NR2C2_disc2* 4.0 (33.4)
 STAT_known2 116.3 (131.0)	 STAT_disc1* 2.4 (56.1)

Table S3 (continued)

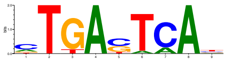
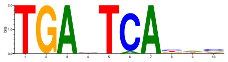
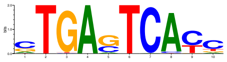
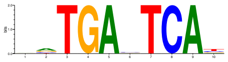
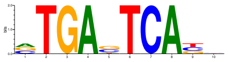
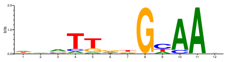
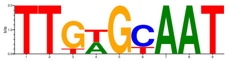
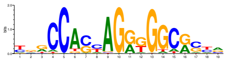
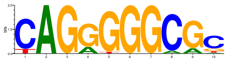
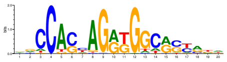
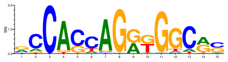
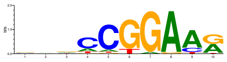
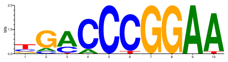
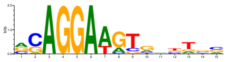
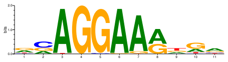
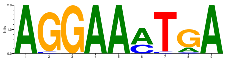
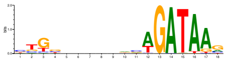
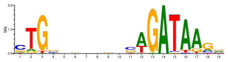
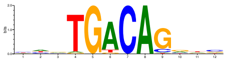
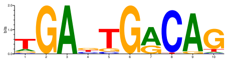
Factor	Matching discovered motifs		
 AP1_known5 (TRE)	 GATA_disc2 (32)	 SMARC_disc1 (34)	 STAT_disc2 (46)
	 TCF7L2_disc1 (35)		
 CEBPB_known7	 STAT_disc4* (47)		
 CTCF_known1	 CTCFL_disc1* (64)	 RAD21_disc1* (62)	 SMC3_disc1 (63)
 ETS_known7	 NR2C2_disc1 (51)		
 ETS_known4	 GATA_disc3 (49)	 MEF2_disc2 (50)	
 GATA_known14	 TAL1_disc1 (66, 67)		
 MEIS1_1	 PBX3_disc2 (65)		

Table S4: Selected shared motifs with literature support. Shown are the motifs that match a known motif for the indicated factor along with relevant citations. Details in the text. * indicates motif is reverse complemented for comparison purposes.

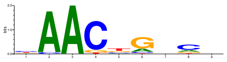
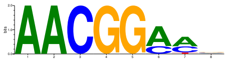
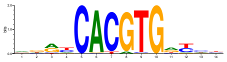
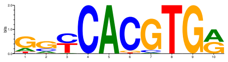
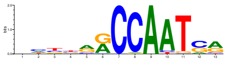
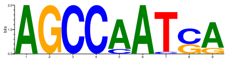
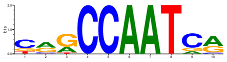
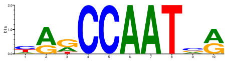
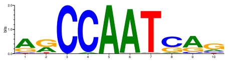
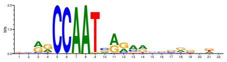
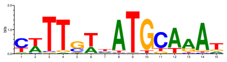
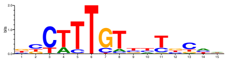
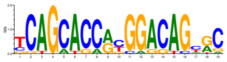
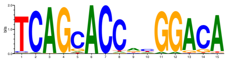
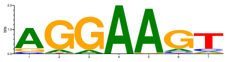
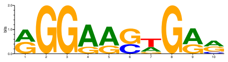
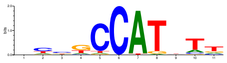
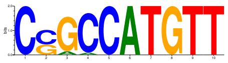
Factor	Matching discovered motifs		
 MYB_4	 ETS_disc8 (54)		
 MYC_known3	 SIN3A_disc2 (38)		
 NFY_known5	 CEBPB_disc2 (43)	 E2F_disc4* (45)	 IRF_disc1 (40)
	 RFX5_disc2 (42)	 SP1_disc1* (44)	
 POU5F1_known3	 NANOG_disc2* (57, 58)		
 REST_known3	 SIN3A_disc1* (37)		
 SPI1_known2	 IRF_disc5 (41)		
 YY1_known5	 THAP1_disc2 (55, 56)		

Table S4 (continued)

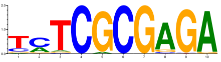
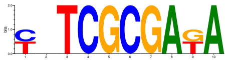
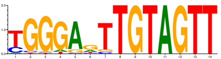
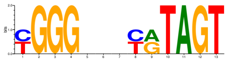
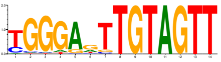
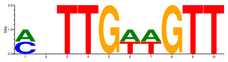
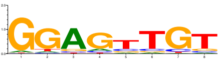
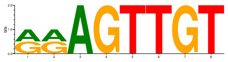
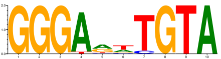
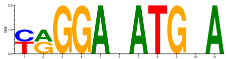
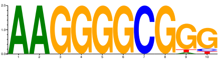
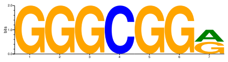
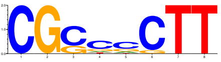
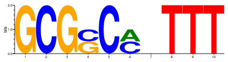
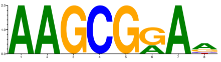
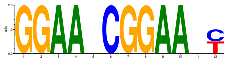
Discovered motif	Motif from (85)
 BRCA1_disc1 (Novel1)	 M8
 ETS_disc1 (Novel2)	 M4
 ETS_disc1 (Novel2)	 M108/FAC1
 ETS_disc5 (Novel2)	 M129
 SIX5_disc2 (Novel2)	 M75/STAT1
 SP2_disc3 (Novel3)	 M6/SP1
 ZBTB7A_disc2 (Novel3)	 M164
 TATA_disc5 (Novel6)	 M21

Table S5: Novel motifs matching those found using systematic conservation measures in four mammals by Xie et al. (85). One representative motif is shown for each putative novel motif when multiple match the same motif. Each motif from (85) is named using the provided identifier and annotation.

References

8. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**, 57–74.

10. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., and Snyder, M. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature,* **489**, 91–100.

11. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research,* **31**, 374–378.

12. Sandelin, A., Alkema, W., Engstrm, P., Wasserman, W. W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research,* **32**, D91–94.

13. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology,* **24**, 1429–1435.

14. Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009) Diversity and Complexity in DNA Recognition by Transcription Factors. *Science,* **324**, 1720–1723.

15. Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Pea-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., Khalid, F., Zhang, W., Newburger, D.,

Jaeger, S. A., Morris, Q. D., Bulyk, M. L., and Hughes, T. R. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell,* **133**, 1266–1276.

16. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell,* **152**, 327–339.

17. Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000) Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology,* **296**, 1205–1214.

18. Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology,* **20**, 835–9.

19. Bailey, T. L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology,* **2**, 28–36.

20. Pavesi, G., Mauri, G., and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics,* **17**, S207–S214.

21. Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods,* **4**, 563–565.

32. Kawana, M., Lee, M. E., Quertermous, E. E., and Quertermous, T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Molecular and Cellular Biology,* **15**, 4225–4231.

34. Ito, T., Yamauchi, M., Nishina, M., Yamamichi, N., Mizutani, T., Ui, M., Murakami, M., and Iba, H. (2001) Identification of SWI.SNF complex subunit BAF60a as a determinant of the transactivation potential of Fos/Jun dimers. *The Journal of Biological Chemistry,* **276**, 2852–2857.

35. Nateri, A. S., Spencer-Dene, B., and Behrens, A. (2005) Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature,* **437**, 281–285.

37. Huang, Y., Myers, S. J., and Dingledine, R. (1999) Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nature Neuroscience,* **2**, 867–872.

38. Nascimento, E. M., Cox, C. L., Macarthur, S., Hussain, S., Trotter, M., Blanco, S., Suraj, M., Nichols, J., Kbler, B., Benitah, S. A., Hendrich, B., Odom, D. T., and Frye, M. (2011) The opposing transcriptional functions of Sin3a and c-Myc are required to maintain tissue homeostasis. *Nature Cell Biology,* **13**, 1395–1405.

40. Li-Weber, M., Davydov, I., Krafft, H., and Krammer, P. (1994) The role of NF-Y and IRF-2 in the regulation of human IL-4 gene expression. *The Journal of Immunology,* **153**, 4122 –4133.

41. Scott, E., Simon, M., Anastasi, J., and Singh, H. (1994) Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science,* **265**, 1573 –1577.

42. Villard, J., Peretti, M., Masternak, K., Barras, E., Caretti, G., Mantovani, R., and Reith, W. (2000) A Functionally Essential Domain of RFX5 Mediates Activation of Major Histocompatibility Complex Class II Promoters by Promoting Cooperative Binding between RFX and NF-Y. *Molecular and Cellular Biology,* **20**, 3364 –3376.

43. Yu, L., Wu, Q., Yang, C. P., and Horwitz, S. B. (1995) Coordination of transcription factors, NF-Y and C/EBP beta, in the regulation of the mdr1b promoter. *Cell Growth & Differentiation: The Molecular Biology Journal of the American Association for Cancer Research,* **6**, 1505–1512.

44. Roder, K., Wolf, S., Larkin, K., and Schweizer, M. (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene,* **234**, 61–69.

45. Caretti, G., Salsi, V., Vecchi, C., Imbriano, C., and Mantovani, R. (2003) Dynamic Recruitment of NF-Y and Histone Acetyltransferases on Cell-cycle Promoters. *Journal of Biological Chemistry,* **278**, 30435 –30440.

46. Ivanov, V. N., Bhoumik, A., Krasilnikov, M., Raz, R., Owen-Schaub, L. B., Levy, D., Horvath, C. M., and Ronai, Z. (2001) Cooperation between STAT3 and c-Jun Suppresses Fas Transcription. *Molecular Cell,* **7**, 517–528.

47. Choi, S., Cho, Y., Kim, H., and Park, J. (2007) ROS mediate the hypoxic repression of the hepcidin gene by inhibiting C/EBPalpha and STAT-3. *Biochemical and Biophysical Research Communications,* **356**, 312–317.

49. Rothbcher, U., Bertrand, V., Lamy, C., and Lemaire, P. (2007) A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development,* **134**, 4023–4032.

50. Taylor, J. M., Dupont-Versteegden, E. E., Davies, J. D., Hassell, J. A., Houl, J. D., Gurley, C. M., and Peterson, C. A. (1997) A role for the ETS domain transcription factor PEA3 in myogenic differentiation. *Molecular and Cellular Biology,* **17**, 5550–5558.

51. O'Geen, H., Lin, Y., Xu, X., Echipare, L., Komashko, V. M., He, D., Frietze, S., Tanabe, O., Shi, L., Sartor, M. A., Engel, J. D., and Farnham, P. J. (2010) Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics,* **11**, 689.

54. Dudek, H., Tantravahi, R. V., Rao, V. N., Reddy, E. S., and Reddy, E. P. (1992) Myb and Ets proteins cooperate in transcriptional activation of the mim-1 promoter. *Proceedings of the National Academy of Sciences,* **89**, 1291 –1295.

55. Mazars, R., Gonzalez-de-Peredo, A., Cayrol, C., Lavigne, A., Vogel, J. L., Ortega, N., Lacroix, C., Gautier, V., Huet, G., Ray, A., Monsarrat, B., Kristie, T. M., and Girard, J. (2010) The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between DYT6 and DYT3 dystonias. *The Journal of Biological Chemistry,* **285**, 13364–13371.

56. Yu, H., Mashtalir, N., Daou, S., Hammond-Martel, I., Ross, J., Sui, G., Hart, G. W., Rauscher, Frank J, r., Drobetsky, E., Milot, E., Shi, Y., and Affar, E. B. (2010) The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. *Molecular and Cellular Biology,* **30**, 5071–5085.

57. Looijenga, L. H. J., Stoop, H., de Leeuw, H. P. J. C., de Gouveia Brazao, C. A., Gillis, A. J. M., van Roozendaal, K. E. P., van Zoelen, E. J. J., Weber, R. F. A., Wolffenbuttel, K. P., van Dekken, H., Honecker, F., Bokemeyer, C., Perlman, E. J., Schneider, D. T., Kononen, J., Sauter, G., and Oosterhuis, J. W. (2003) POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. *Cancer Research,* **63**, 2244–2250.

58. Loh, Y., Wu, Q., Chew, J., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K., Sung, K. W., Lee, C. W. H., Zhao, X., Chiu, K., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C., Ruan, Y., Lim, B., and Ng, H. (2006) The Oct4 and Nanog

transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics,* **38**, 431–440.

62. Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., and Peters, J. (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature,* **451**, 796–801.

63. Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A., and Krumm, A. (2008) CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America,* **105**, 8309–8314.

64. Jelinic, P., Stehle, J., and Shaw, P. (2006) The Testis-Specific Factor CTCFL Cooperates with the Protein Methyltransferase PRMT7 in H19 Imprinting Control Region Methylation. *PLoS Biol,* **4**, e355.

65. Bischof, L. J., Kagawa, N., Moskow, J. J., Takahashi, Y., Iwamatsu, A., Buchberg, A. M., and Waterman, M. R. (1998) Members of the Meis1 and Pbx Homeodomain Protein Families Cooperatively Bind a cAMP-responsive Sequence (CRS1) from BovineCYP17. *Journal of Biological Chemistry,* **273**, 7941 –7948.

66. Kappel, A., Schlaeger, T. M., Flamme, I., Orkin, S. H., Risau, W., and Breier, G. (2000) Role of SCL/Tal-1, GATA, and ets transcription factor binding sites for the regulation of flk-1 expression during murine vascular development. *Blood,* **96**, 3078–3085.

67. Mouthon, M. A., Bernard, O., Mitjavila, M. T., Romeo, P. H., Vainchenker, W., and Mathieu-Mahul, D. (1993) Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood,* **81**, 647–655.

85. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. *Nature,* **434**, 338–345.

97. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alfoldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M.,

Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., and Kellis, M. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature,* **478**, 476–482.

101. MacArthur, S., Li, X., Li, J., Brown, J., Chu, H. C., Zeng, L., Grondona, B., Hechmer, A., Simirenko, L., Keranen, S., Knowles, D., Stapleton, M., Bickel, P., Biggin, M., and Eisen, M. (2009) Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology,* **10**, R80.

102. Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments.. *Nucleic Acids Research,* **24**, 3836–3845.

103. Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., and Bruford, E. A. (2013) Genenames.org: the HGNC resources in 2013. *Nucleic acids research,* **41**, D545–552.

104. Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotech,* **26**, 1351–1359.

105. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002) The Human Genome Browser at UCSC. *Genome Research,* **12**, 996 –1006.

106. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guig, R., and Hubbard, T. J. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome research,* **22**, 1760–1774.

107. Touzet, H. and Varre, J. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for Molecular Biology,* **2**, 15.

108. Wilson, E. B. (1927) Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association,* **22**, 209–212.

109. Mahony, S., Auron, P. E., and Benos, P. V. (2007) DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Comput Biol,* **3**, e61.

110. Sandelin, A. and Wasserman, W. W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of molecular biology,* **338**, 207–215.