Supplementary Material

On the Detection and Refinement of Transcription Factor Binding Sites Using ChIP-Seq Data

Ming Hu, Jindan Yu, Jeremy M. G. Taylor, Arul M. Chinnaiyan, and Zhaohui Qin

1. URLs of Data Used in This Study:

ChIP-Seq:
NRSF: http://www.illumina.com/downloads/Illumina_ChIPSeq_Demo_Data_Johnson_Science_2007.zip
STAT1: http://www.bcgsc.ca/data/chipseq
CTCF: http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx
ER: submitted to the GEO database; the accession number is GSE19013.

ChIP-chip:
NRSF: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds&term=GSE8489[Accession]&cmd=search
STAT1: http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE2714
CTCF: http://licr-renlab.ucsd.edu/download.html
ER:
http://research4.dfci.harvard.edu/brownlab//datasets/index.php?dir=ER_MCF7_whole_human_genome/

2. HPeak Software:

HPeak is a hidden Markov model (HMM)-based algorithm for analyzing ChIP-Seq data. The goal of HPeak is to partition the genome into segments that are either ChIP-enriched or non-enriched such that the enriched portion of the genome is much more likely to harbor protein-DNA interaction sites. The input data is a collection of sequencing reads that have been aligned to the reference genome uniquely. HPeak first partitions the entire genome into small bins of fixed length (e.g., 25 bps) and evaluates the distribution of ChIP DNA fragments in these bins throughout the genome. Next HPeak applies a two-state HMM on the sequencing depth profile to identify stretches of ChIP-enriched bins from the background. HPeak uses two different probability distributions, the generalized Poisson (GP) distribution (1) and the zero inflated Poisson (ZIP) distribution (2) to model the numbers of sequencing reads that overlap with ChIP-

enriched and non-enriched bins respectively. Both these distributions are modified from the standard Poisson distribution to fit data where there is serious over or under dispersion or there are a large proportion of extra zeros, cases we often observe in genome-wide sequencing read coverage profiles. Using a user-specified posterior probability threshold, HPeak then identifies stretches of bins from the HMM that show significant enrichment of sequencing read counts. Each set of bins is defined as a peak. In addition to its genomic location and the length of the peak, HPeak also reports the location of the highest sequencing depth within the peak, the actual maximum sequencing depth at that location, and the posterior probability of these bins being ChIP-enriched. Because such probability reflects the significance of these peaks, one can rank all peaks predicted by HPeak using these probabilities. HPeak software is freely available from the website http://www.sph.umich.edu/csg/qin/HPeak/.

3. Estrogen Receptor ChIP-Seq Experiment on MCF7 Cells

MCF-7 cells were grown in RPMI media supplemented with 10% FBS to 50% confluence. The cells were then hormone-starved for three days prior to the treatment with 10 nM β-estradiol or vehicle control for 45 minutes. The cells were then harvested for ChIP analysis as previously described using an antibody against the estrogen receptor (ER)-alpha (sc-543x, Santa Cruz Biotechnology Inc, Santa Cruz, CA) or the IgG control. Briefly, cultured cells near 90% confluence were crosslinked with 1% formaldehyde for 10 minutes and the crosslinking was inactivated by 0.125 M glycine for 5 minutes at room temperature (RT). The cells were then rinsed with cold 1X PBS twice and scraped off in 1X PBS + protease inhibitor (PI). Cells were pelleted and resuspended in cell lysis buffer plus PI for 10 minutes. Nuclei pellets were spun at 5,000 rpm for 5 minutes, resuspended in nuclear lysis buffer, and then incubated for 10 minutes. Chromatin was sonicated to an average length of 500 bp with an Ultrasonic Processor Sonicator 3000 (Misonix Inc, Farmingdale, NY) and then centrifuged at 14,000 rpm for 10 minutes to remove the debris. Supernatants containing chromatin fragments were incubated with agarose/protein A or G beads (Millipore. Billerica, MA) for 15 minutes and centrifuged at 5,000 rpm for 5 minutes to remove the nonspecific binding. To immunoprecipitate protein/chromatin complexes, the supernatants were incubated with 3-5 µg of antibody or IgG overnight, then added 50 µl of agarose/protein A or G beads and incubated for another hour. Beads were washed twice with 1X dialysis buffer and four times with IP buffer. The antibody/protein/DNA complexes were eluted with 150 µL IP elution buffer twice. To reverse the crosslinks, the complexes were incubated in elution buffer plus 10 µg RNase A and 0.3 M NaCl at 67°C for four hours. DNA/proteins were precipitated with ethanol, air-dried, and dissolved in 100 µL of

TE. Proteins were then digested by proteinase K at 45°C for one hour and DNA was purified with a QIAGEN PCR purification column and eluted with 30 µL EB buffer.

The ChIP-enriched DNA was evaluated for significant enrichment of positive control genes and then subjected to ChIP-Seq sample preparation following the manufacturer's protocols (Illumina Inc, San Diego, CA). Briefly, the ends of ChIP-enriched DNA or control DNA (~10 ng) was first repaired by T4 DNA polymerase, T4 PNK, and Klenow DNA polymerase at 20°C for 30 minutes. An "A" base was added to the 3' end of the blunt phosphorylated DNA fragments using Klenow exo at 37°C for 30 minutes. Adapters were then ligated to the ends of the DNA fragments by DNA ligase at RT for 15 minutes. DNA fragments were separated on 2% gel at 100V for 1 hour, the 200±25bp band was excised from the gel, and the DNA was extracted by QIAGEN gel extraction kit. Gel-extracted DNA was amplified by PCR reaction for 16 cycles and quality assured using Bioanalyzer (Agilent Technologies, Santa Clara, CA). ChIP-Sequencing was performed using the Illumina Genome Analyzer according to standard manufacturer's procedures. The raw sequencing image data were analyzed by the Illumina analysis pipeline, aligned to the unmasked human reference genome (NCBI v36, hg18) using Eland software (Illumina Inc, San Diego, CA) to generate sequence reads of 35 bps.

4. Simulation scheme for studying intra-motif dependency

As described in the Material and Methods Section of the manuscript, for a pair of positions within the motif, we use the Hamming distance between two sets of estimated dinucleotide frequencies based on two competing probability models (16-component multinomial distribution or the product of two four-component multinomial distributions) to gauge whether the two positions are dependent. To select a reasonable cutoff, we conducted a simulation study to estimate the null distribution for such Hamming distances. We considered five levels of nucleotide conservation, with information content ranging from 0.29 to 1.76. There are in total 15 different combinations of these information contents. For each combination, we specify two four-component multinomial distributions that match the two information content levels. One thousand nucleotides were drawn from each of the two multinomial distributions independently. We choose the large number to reflect the fact that typically large amounts of motifs were identified from ChIP-Seq experiments. The Hamming distances were calculated using formula (4) in the manuscript:

$$d_{12} = \sum_{x=1}^{4}\sum_{y=1}^{4}\left|\hat{\eta}_{xy}(r_1, r_2) - \hat{\eta}_x(r_1)\hat{\eta}_y(r_2)\right|.$$

We simulated one million position pairs using the above procedure in order to obtain an accurate null distribution of the background Hamming distance. The histograms of these Hamming distances were shown in Figure S8. From these plots we found that strong dependency (large Hamming distance) tends to occur between a pair of positions in which each position itself is weakly conserved.

5. Multi-nucleotide logo plots

HMS is able to identify the position pairs that show intra-motif dependency. To visualize the multi-nucleotide preference at the dependent positions, we extended the traditional logo plot (3) to multi-nucleotide logo plot. The overall height is the information content which can be calculated using the following two formulas: $4 + \sum_{i\in\{A,C,G,T\}}\sum_{j\in\{A,C,G,T\}} p_{ij}\log_2(p_{ij})$ for dinucletides and $6 + \sum_{i\in\{A,C,G,T\}}\sum_{j\in\{A,C,G,T\}}\sum_{k\in\{A,C,G,T\}} p_{ijk}\log_2(p_{ijk})$ for trinucleotides. The multi-nucleotides are sorted by their frequencies from top to bottom, and the height of each multi-nucleotide is proportional to its frequency. The top multi-nucleotide is the most enriched one at the dependent positions. The R code we wrote to draw these extended logo plots was modified from the SeqLogo program written by Bembom (3). The R code is freely available at http://www.sph.umich.edu/csg/qin/HMS/.

Table S1A. Four motif models for two motif widths and two motif strengths used in the simulation study.

Motif Consensus (Width)	Motif information content	
	Strong	**Weak**
GACTACCA (Width = 8)	1.42	0.93
AGGATCTAATGATCCT (Width =16)	1.42	0.93

Table S1B. Two motif abundances scheme used in the simulation study.

Expected copies of motif segments	Motif abundance	
	Abundant	**Sparse**
Top 25% sequences	0.9	0.6
25% - 50% sequences	0.7	0.4
50% - 75% sequences	0.5	0.2
Last 25% sequences	0.3	0
Total expected motif segments	1,800	900

Table S2A. Performance comparison on simulated data assuming all positions within the motif are independent.

Simulation setting[1]			Times_found[2]				Difference compared to HMS[3]		
Width	Information content	Abundance	HMS	HMS_uniform	MEME	MDscan	HMS_uniform	MEME	MDscan
8	Strong	Abundant	100	100	100	100	0.01**	0.00	0.11**
8	Strong	Sparse	100	100	100	100	0.00**	0.00	0.09**
8	Weak	Abundant	100	100	71	33	0.00	0.07**	0.11**
8	Weak	Sparse	100	100	78	96	0.01	0.09**	0.02
16	Strong	Abundant	100	100	100	100	0.00	0.02**	0.12**
16	Strong	Sparse	100	100	100	100	0.00**	0.02**	0.10**
16	Weak	Abundant	100	100	100	54	0.00*	0.02**	0.21**
16	Weak	Sparse	100	100	100	91	0.00*	0.02**	0.20**

Table S2B. Performance comparison on simulated data assuming some positions within the motif are dependent.

Simulation setting[1]			Times_found[2]				Difference compared to HMS[3]		
Width	Information content	Abundance	HMS	HMS_uniform	MEME	MDscan	HMS_uniform	MEME	MDscan
8	Strong	Abundant	98	98	63	96	0.00	0.10**	0.11**
8	Strong	Sparse	100	100	85	93	0.00	0.09**	0.10**
8	Weak	Abundant	92	89	0	2	-0.01#	NA	0.09#
8	Weak	Sparse	73	69	0	61	0.03	NA	0.09**
16	Strong	Abundant	100	100	100	100	0.00	0.22**	0.28**
16	Strong	Sparse	100	100	100	100	0.00	0.21**	0.26**
16	Weak	Abundant	100	100	100	40	0.00	0.21**	0.31**
16	Weak	Sparse	100	100	100	83	0.00	0.20**	0.30**

[1]Each simulation setting is a combination of motif width, information content and motif abundance. The scheme is similar to Liu et al, (2001) (4) and described in Table S1.

[2]"Times found" indicates among the 100 simulated dataset, how many times the correct motif is identified by the motif-finding algorithm.

[3]Difference referrers to the difference between two average Hamming distances h and h_{HMS} in which h measures average Hamming distance between probability matrix $\boldsymbol{\Theta}$ and its prediction denoted as $\hat{\boldsymbol{\Theta}}$: $h = \frac{1}{w}\sum_{i=1}^{4}\sum_{j=1}^{w}\left|\theta_{ij} - \hat{\theta}_{ij}\right|$ (Formula (11) in the manuscript). h_{HMS} measures average Hamming distance between probability matrix $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Theta}}$ predicted by HMS method. We use * to indicate a p-value in paired t-test between 0.01 and 0.05 and ** to indicate a p-value in paired t-test less than 0.01. [#] indicates that p-value is not available due to lack of sample size (in simulation setting width 8, weak information content and abundant abundance, MDscan can detect correct motif only in two out 100 simulated data sets.)

Table S3. The joint distribution of dinucleotides in two dependent positions. The probabilities in this multinomial distribution are taken from the one predicted for position pair (1,2) in the E2F motif in Zhou and Liu (5) (Figure 2(b) in their paper).

First Base	**Second Base**			
	A	**C**	**G**	**T**
A	0	0	0.19	0
C	0	0	0.16	0.06
G	0	0.09	0	0
T	0	0.44	0.06	0

Table S4A. Information on the four real ChIP-Seq datasets

TF	Cell type	Antibody	# of peaks	Coverage	Reference
NRSF	Jurkat T cell	Monoclonal antibody 12C11	4,982	1.4 MB	Johnson et al. (2007)
STAT1	HeLa S3 cell	Rabbit polyclonal antibody	27,470	8.1 MB	Robertson et al. (2007)
CTCF	CD4+ T cell	Upstate 07-729	22,159	7.4 MB	Barski et al. (2007)
ER	MCF7 cell	ER α (HC-20)	10,072	2.5 MB	(current manuscript)

Table S4B. Information on the four real ChIP-chip datasets

TF	Cell type	Antibody	# of peaks	Coverage	Reference
NRSF	Jurkat T cell	Monoclonal antibody	8,819	12.2 MB	Johnson et al. (2007)
STAT1	HeLa S3 cell	α p91 (C-24) rabbit polyclonal antibody	3,701	4.7 MB	Euskirchen et al. (2007)
CTCF	IMR90 and U937 cell	Mixture of 9 monoclonal antibodies	13,804	12.1 MB	Kim et al. (2005)
ER	MCF7 cell	ER α (HC-20)	10,901	11.1 MB	Carroll et al. (2005)

Table S5. Dependent positions identified in four motifs from ChIP-Seq data.

Motif	Top dependent pairs	Hamming distance
NRSF width = 21 bp	(18-19)	0.3308
Dependent positions:	(3-4)	0.3297
[(2-3)(3-4)] [(18-19)(19-20)]	(19-20)	0.3196
	(2-3)	0.2629
CTCF width = 24 bp	(14-15)	0.4635
Dependent positions:	(16-17)	0.3358
[(13-14)(14-15)] [(16-17)(17-18)]	(17-18)	0.2720
	(13-14)	0.2664
STAT11 width = 19 bp	(7-8)	0.3052
Dependent positions:	(6-7)	0.2853
[(6-7)(7-8)] (13-14)	(13-14)	0.2008
ER width = 19 bp	(11-12)	0.2562
Dependent positions:	(10-11)	0.2553
[(2-3)(3-4)] [(10-11)(11-12)] (18-19)	(18-19)	0.2327
	(3-4)	0.2260
	(2-3)	0.2070

Table S6. The probability mass function of 16-component multinomial distribution (dependent), and the probability mass function of 16-component multinomial distribution (independent: the outer product of the probabilities of two independent four-component multinomial distributions) in 16 dependent position pairs identified from four motifs. The dinucleotides with absolute difference between the two probabilities above 0.05 are highlighted in bold.

A.

NRSF	Position 19				Marginal distribution
Position 18	A	C	G	T	
A	0.02 (0.03)	0.00 (0.00)	0.09 (0.06)	0.02 (0.04)	0.13
C	**0.18 (0.13)**	0.01 (0.02)	**0.17 (0.26)**	0.23 (0.19)	0.60
G	0.02 (0.03)	0.00 (0.00)	0.08 (0.05)	0.02 (0.04)	0.12
T	0.02 (0.03)	0.00 (0.00)	0.09 (0.07)	0.05 (0.05)	0.15
Marginal distribution	0.22	0.02	0.44	0.32	1

B.

NRSF	Position 4				Marginal distribution
Position 3	A	C	G	T	
A	0.03 (0.06)	0.00 (0.00)	0.03 (0.01)	0.00 (0.00)	0.07
C	**0.73 (0.65)**	0.01 (0.03)	**0.02 (0.08)**	0.01 (0.02)	0.78
G	0.02 (0.05)	0.01 (0.00)	0.02 (0.01)	0.01 (0.00)	0.06
T	0.05 (0.08)	0.01 (0.00)	0.03 (0.01)	0.00 (0.00)	0.09
Marginal distribution	0.84	0.03	0.11	0.02	1

C.

NRSF	Position 20				Marginal distribution
Position 19	A	C	G	T	
A	0.04 (0.03)	0.10 (0.14)	0.07 (0.04)	0.02 (0.02)	0.23
C	0.01 (0.00)	0.01 (0.02)	0.00 (0.00)	0.00 (0.00)	0.02
G	0.07 (0.06)	0.22 (0.26)	0.1 (0.08)	0.04 (0.03)	0.43
T	0.02 (0.04)	**0.27 (0.19)**	0.02 (0.06)	0.01 (0.02)	0.31
Marginal distribution	0.13	0.61	0.18	0.07	1

D.

NRSF	Position 3				Marginal distribution
Position 2	A	C	G	T	
A	0.01 (0.00)	0.03 (0.04)	0.01 (0.00)	0.01 (0.01)	0.05
C	0.03 (0.02)	0.13 (0.16)	0.01 (0.01)	0.04 (0.02)	0.21
G	0.01 (0.01)	0.07 (0.09)	0.01 (0.01)	0.02 (0.01)	0.12
T	0.02 (0.04)	**0.54 (0.48)**	0.03 (0.04)	0.03 (0.06)	0.62
Marginal distribution	0.07	0.77	0.06	0.10	1

E.

CTCF	Position 15				Marginal distribution
Position 14	A	C	G	T	
A	0.03 (0.04)	**0.14 (0.25)**	**0.28 (0.16)**	0.03 (0.03)	0.48
C	0.00 (0.00)	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02
G	0.05 (0.04)	**0.34 (0.25)**	**0.06 (0.16)**	0.03 (0.03)	0.48
T	0.00 (0.00)	0.02 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02
Marginal distribution	0.08	0.52	0.34	0.06	1

F.

CTCF	Position 17				Marginal distribution
Position 16	A	C	G	T	
A	0.04 (0.05)	0.03 (0.03)	0.04 (0.03)	0.01 (0.01)	0.12
C	**0.21 (0.15)**	0.08 (0.09)	**0.04 (0.11)**	0.03 (0.03)	0.38
G	0.02 (0.03)	0.04 (0.02)	0.02 (0.02)	0.01 (0.01)	0.08
T	**0.12 (0.17)**	0.09 (0.10)	**0.20 (0.13)**	0.02 (0.03)	0.43
Marginal distribution	0.40	0.24	0.29	0.08	1

G.

CTCF	Position 18				Marginal distribution
Position 17	A	C	G	T	
A	0.07 (0.07)	0.06 (0.08)	**0.16 (0.11)**	0.11 (0.14)	0.40
C	0.05 (0.04)	0.04 (0.05)	0.04 (0.07)	0.10 (0.08)	0.24
G	0.03 (0.05)	0.08 (0.06)	0.05 (0.08)	0.13 (0.10)	0.29
T	0.01 (0.01)	0.01 (0.01)	0.03 (0.02)	0.02 (0.03)	0.07
Marginal distribution	0.17	0.20	0.28	0.35	1

H.

CTCF	Position 14				Marginal distribution
Position 13	A	C	G	T	
A	0.02 (0.05)	0.00 (0.00)	0.09 (0.05)	0.00 (0.00)	0.10
C	**0.45 (0.38)**	0.01 (0.01)	**0.32 (0.38)**	0.02 (0.02)	0.79
G	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00
T	0.01 (0.04)	0.00 (0.00)	0.07 (0.04)	0.00 (0.00)	0.08
Marginal distribution	0.47	0.01	0.47	0.02	1

I.

STAT1	Position 8				Marginal distribution
Position 7	A	C	G	T	
A	0.06 (0.08)	0.00 (0.00)	0.21 (0.19)	0.01 (0.02)	0.29
C	**0.13 (0.06)**	0.00 (0.00)	**0.06 (0.13)**	0.01 (0.01)	0.20
G	0.04 (0.05)	0.00 (0.00)	0.13 (0.12)	0.01 (0.01)	0.18
T	0.06 (0.10)	0.00 (0.00)	0.25 (0.22)	0.02 (0.02)	0.34
Marginal distribution	0.29	0.00	0.66	0.06	1

J.

STAT1	Position 7				Marginal distribution
Position 6	A	C	G	T	
A	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.02)	0.05
C	0.21 (0.19)	0.15 (0.14)	**0.05 (0.12)**	0.25 (0.22)	0.67
G	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00
T	0.05 (0.08)	0.04 (0.06)	**0.12 (0.05)**	0.07 (0.09)	0.28
Marginal distribution	0.28	0.21	0.18	0.33	1

K.

STAT1	Position 14				Marginal distribution
Position 13	A	C	G	T	
A	0.04 (0.04)	0.06 (0.06)	0.08 (0.07)	0.03 (0.04)	0.21
C	0.06 (0.04)	0.07 (0.06)	**0.02 (0.07)**	0.06 (0.04)	0.21
G	0.04 (0.03)	0.04 (0.05)	0.05 (0.05)	0.03 (0.03)	0.16
T	0.06 (0.08)	0.11 (0.12)	0.18 (0.14)	0.06 (0.07)	0.41
Marginal distribution	0.19	0.29	0.33	0.18	1

L.

ER	Position 12				Marginal distribution
Position 11	A	C	G	T	
A	0.04 (0.04)	0.05 (0.06)	0.10 (0.08)	0.02 (0.03)	0.21
C	0.07 (0.05)	0.11 (0.09)	**0.05 (0.11)**	0.07 (0.04)	0.29
G	0.05 (0.05)	0.08 (0.08)	0.12 (0.10)	0.03 (0.04)	0.27
T	0.02 (0.04)	0.06 (0.07)	0.10 (0.08)	0.03 (0.03)	0.22
Marginal distribution	0.18	0.30	0.37	0.14	1

M.

ER	Position 11				Marginal distribution
Position 10	A	C	G	T	
A	0.03 (0.03)	0.04 (0.04)	0.06 (0.04)	0.02 (0.03)	0.14
C	0.11 (0.08)	0.13 (0.11)	**0.05 (0.11)**	0.10 (0.08)	0.38
G	0.06 (0.06)	0.07 (0.08)	0.11 (0.08)	0.05 (0.06)	0.28
T	0.02 (0.04)	0.05 (0.05)	0.07 (0.05)	0.04 (0.04)	0.18
Marginal distribution	0.21	0.28	0.28	0.21	1

N.

ER	Position 19				Marginal distribution
Position 18	A	C	G	T	
A	0.02 (0.02)	0.01 (0.02)	0.03 (0.02)	0.02 (0.02)	0.08
C	0.11 (0.06)	0.06 (0.06)	**0.04 (0.09)**	0.09 (0.09)	0.30
G	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.01 (0.01)	0.04
T	**0.07 (0.12)**	0.13 (0.12)	0.21 (0.17)	0.16 (0.16)	0.57
Marginal distribution	0.21	0.21	0.29	0.28	1

O.

ER	Position 4				Marginal distribution
Position 3	A	C	G	T	
A	0.17 (0.17)	0.01 (0.01)	0.10 (0.09)	0.02 (0.02)	0.29
C	0.21 (0.17)	0.01 (0.01)	**0.04 (0.09)**	0.03 (0.02)	0.29
G	0.12 (0.12)	0.01 (0.01)	0.06 (0.06)	0.01 (0.01)	0.20
T	0.07 (0.12)	0.01 (0.01)	**0.12 (0.07)**	0.02 (0.02)	0.22
Marginal distribution	0.58	0.04	0.31	0.07	1

P.

ER	Position 3				Marginal distribution
Position 2	A	C	G	T	
A	0.06 (0.06)	0.06 (0.06)	0.06 (0.04)	0.03 (0.05)	0.21
C	0.14 (0.11)	0.11 (0.11)	**0.03 (0.08)**	0.10 (0.08)	0.38
G	0.05 (0.05)	0.05 (0.05)	0.04 (0.03)	0.03 (0.04)	0.17
T	0.04 (0.07)	0.07 (0.07)	0.07 (0.05)	0.05 (0.05)	0.24
Marginal distribution	0.28	0.21	0.18	0.33	1

Table S7. 16 dinucleotide frequencies for positions 3-4 and positions 18-19 in ER motif identified by HMS.

Positions 3 and 4	Dinucleotide frequency	Positions 18 and 19	Dinucleotide frequency
AA	0.1661	TT	0.1628
AC	0.0118	GT	0.0119
AG	0.0977	CT	0.0937
AT	0.0169	AT	0.0173
CA	0.2080	TG	0.2130
CC	0.0149	GG	0.0152
CG	0.0416	CG	0.0403
CT	0.0271	AG	0.0280
GA	0.1228	TC	0.1293
GC	0.0089	GC	0.0090
GG	0.0574	CC	0.0570
GT	0.0128	AC	0.0122
TA	0.0729	TA	0.0712
TC	0.0094	GA	0.0096
TG	0.1152	CA	0.1127
TT	0.0166	AA	0.0168

Table S8. Comparison of motif enrichment among motif patterns identified by different *de novo* motif finding tools as well as known motif patterns described in the literature or stored in the MatBase (Genomatix Software GmBH, Munich, Germany).

Area under the curve[1]	ChIP-Seq[2]				ChIP-chip[3]			
	NRSF	**STAT1**	**CTCF**	**ER**	**NRSF**	**STAT1**	**CTCF**	**ER**
HMS	**258.62****	439.68	**2312.87****	**327.99****	**163.08****	**94.23****	**1102.02****	**1216.94****
HMS_ind	254.73**	388.99	2198.81**	290.30	161.98**	83.75	1001.35**	1035.42**
MEME	242.83	**440.33**	2076.76	281.14	148.76	82.95	908.35	773.30
MDscan	240.16	--	--	--	143.25	--	--	--
Genomatix V$NRSF	210.38	--	--	--	36.38	--	--	--
Genomatix V$STAT01	--	77.39	--	--	--	9.69	--	--
Genomatix V$STAT03	--	191.70	--	--	--	28.96	--	--
Kim07_CTCF	--	--	1225.97	--	--	--	509.62	--
Genomatix V$ER01	--	--	--	35.29	--	--	--	292.43
Genomatix V$ER02	--	--	--	35.05	--	--	--	71.48
Genomatix V$ER03	--	--	--	87.40	--	--	--	245.37

[1]Area under the curve (AUC) in the empirical FDR versus Chi-square test statistics plot (Figure 3B and C, Figures S4-6 B and C). Values in bold indicate the best performance in that column. Five cross-validations were performed on each dataset using each of the four motif finding algorithms. In addition, we conducted cross-validation 100 times and compared the AUCs obtained from two different method using a paired t-test to assess whether the performance difference we observed in statistical significant. We use ** indicates an empirical p-value less than 0.01.

[2]The empirical FDRs for NRSF, STAT1, CTCF and ER ChIP-Seq data sets all range from 0 to 0.2.

[3]The empirical FDRs for NRSF and CTCF ChIP-chip data sets range from 0 to 0.2. The empirical FDRs for STAT1 and ER ChIP-chip data sets range from 0 to 1.

Table S9. Comparison of motif enrichment among the three motif finding tools and known motif patterns stored in the MatBase (Genomatix GmBH, Munich, Germany).[1]

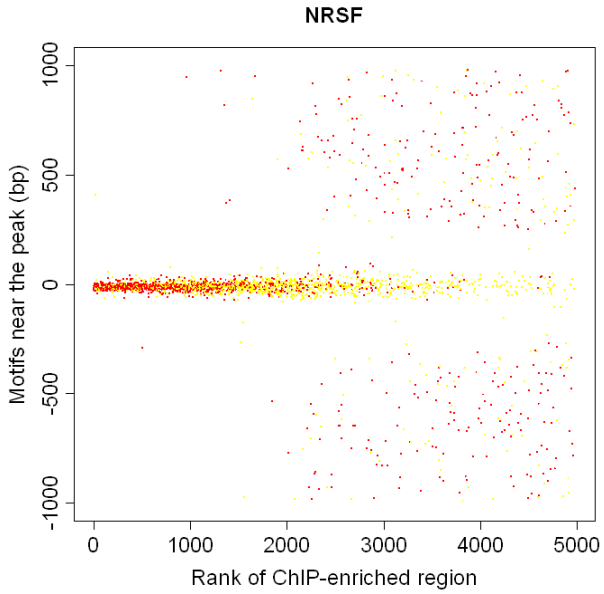
ChIP-Seq	NRSF		STAT1		CTCF		ER	
Empirical FDR[2]	**0.05**	**0.1**	**0.05**	**0.1**	**0.05**	**0.1**	**0.05**	**0.1**
HMS	**49.30**	**51.41**	**10.84**	**25.29**	**79.05**	**84.45**	**29.39**	**38.57**
HMS_ind	48.25	51.09	10.12	22.10	75.91	83.10	26.02	35.69
MEME	46.63	49.55	10.28	24.08	72.43	81.16	24.59	35.32
MDscan	45.21	49.11	--	--	--	--	--	--
Genomatix V$NRSF	41.09	42.52	--	--	--	--	--	--
Genmoatix V$STAT01	--	--	--	--	--	--	--	--
Genomatix V$STAT03	--	--	10.09	12.16	--	--	--	--
Kim07_CTCF	--	--	--	--	56.68	62.01	--	--
Genomatix V$ER01	--	--	--	--	--	--	--	--
Genomatix V$ER02	--	--	--	--	--	--	--	--
Genomatix V$ER03	--	--	--	--	--	--	10.62	14.56

[1]Values in the table are percentages of ChIP-enriched sequences that contain the specific motif pattern. Values in bold indicate the best performance in that column.

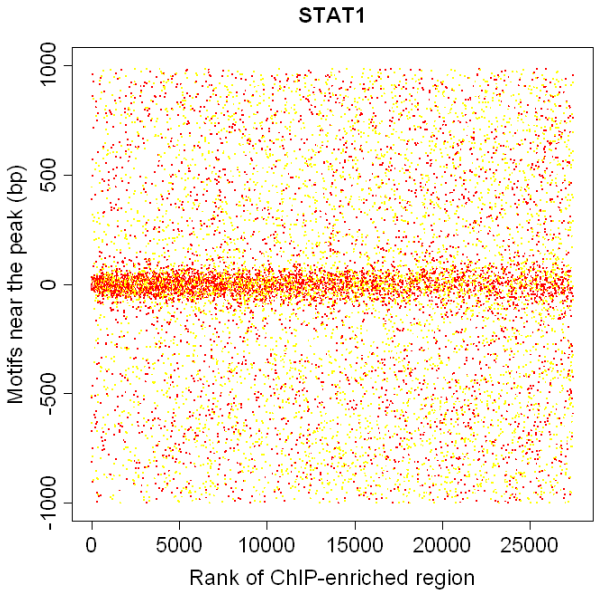
[2]The empirical FDR is estimated by dividing the number of control sequences that contain the motif by the number of testing sequences that contain the motif.

Figure S1. Rank order of all ChIP-enriched regions versus location of motifs. Zero in the y-axis indicates the location of highest sequencing depth in the ChIP-enriched regions obtained from HPeak program (see section 2 of this document). For each position $a_j = l$ within a ChIP-enriched region $\boldsymbol{R}_j$, we calculate a motif score defined as equation (1) in the main text, measuring the similarity between the DNA sequence of length *w* (motif length, assumed known) starts from the current location and the known motif pattern represented by PSWM. Higher scores indicates better match. We record the position with highest motif score for each ChIP-enriched region $\boldsymbol{R}_j$. For each dot, the x-axis represents the rank of ChIP-enriched region (from the highest to the lowest), and the y-axis represents the physical position of the most likely motif location in each ChIP-enriched region. The red dots indicate the motifs with score above the first quantile, and the yellow dots indicate the motifs with score between the first quantile and median. **A. NRSF, B. STAT1, C. CTCF, D. ER.**

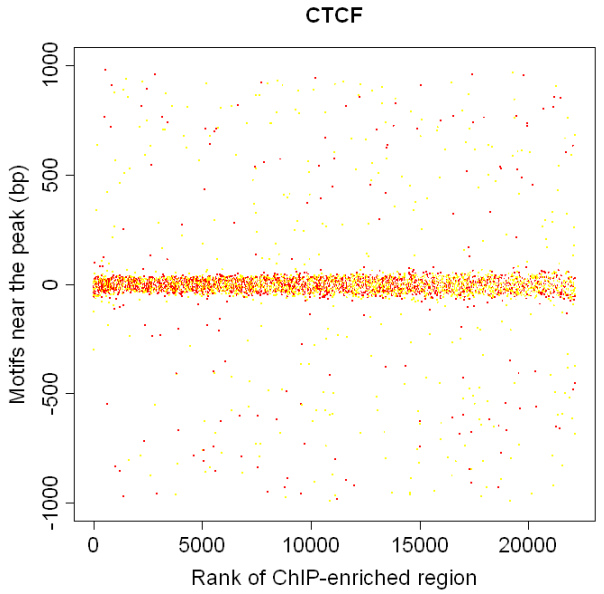
A.


NRSF
Motifs near the peak (bp)
1000
500
0
-500
-1000
0
1000
2000
3000
4000
5000
Rank of ChIP-enriched region

B.


STAT1
Motifs near the peak (bp)
1000
500
0
-500
-1000
0
5000
10000
15000
20000
25000
Rank of ChIP-enriched region

C.


CTCF
Motifs near the peak (bp)
1000
500
0
-500
-1000
0
5000
10000
15000
20000
Rank of ChIP-enriched region

D.

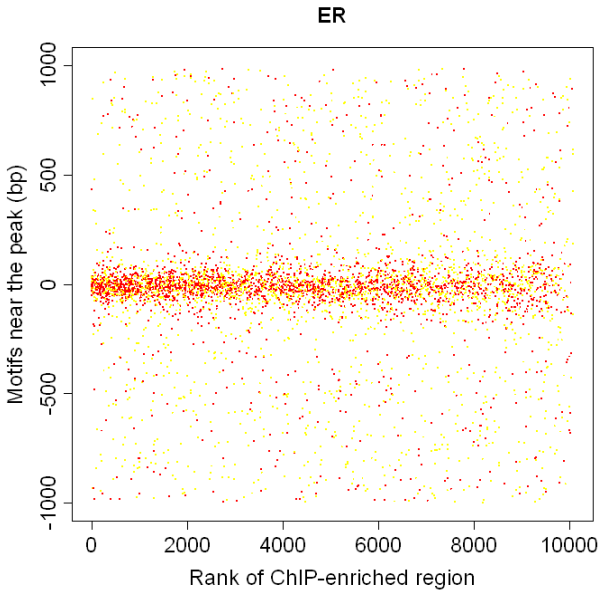

ER
Motifs near the peak (bp)
1000
500
0
-500
-1000
0
2000
4000
6000
8000
10000
Rank of ChIP-enriched region

Figure S2. Illustration of the informative prior distribution of motif start locations. The prior probabilities (solid black line) are proportional to a discretized Student's *t* distribution with three degrees of freedom (with standard error = 1.73) and rescaled such that the prior probabilities form a step function with a fixed step-size (25 bp in this study). The solid red line represents the probability density function of shifted and rescaled Student's *t* distribution with three degrees of freedom.

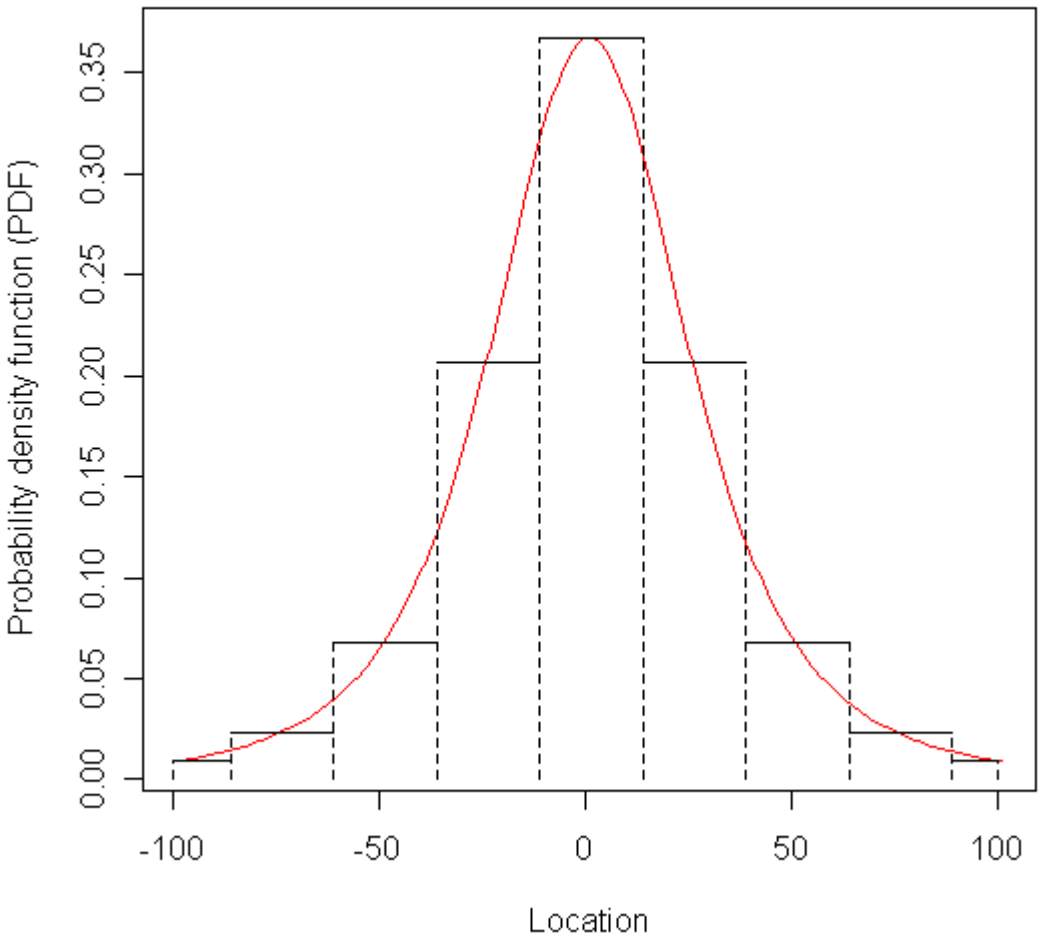
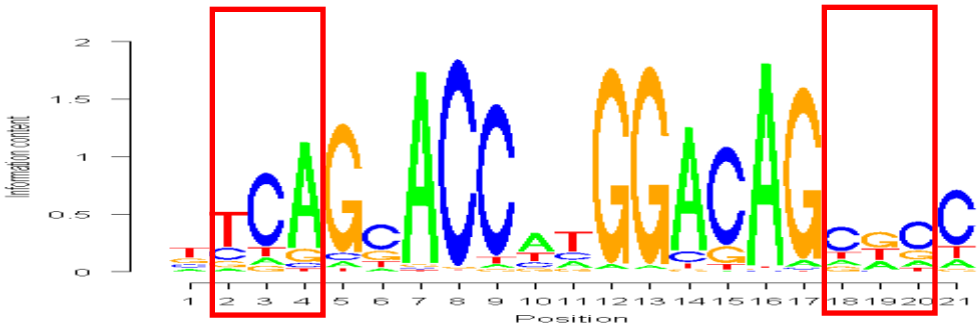
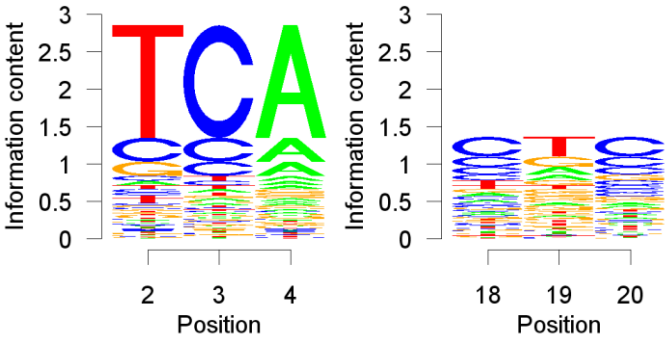
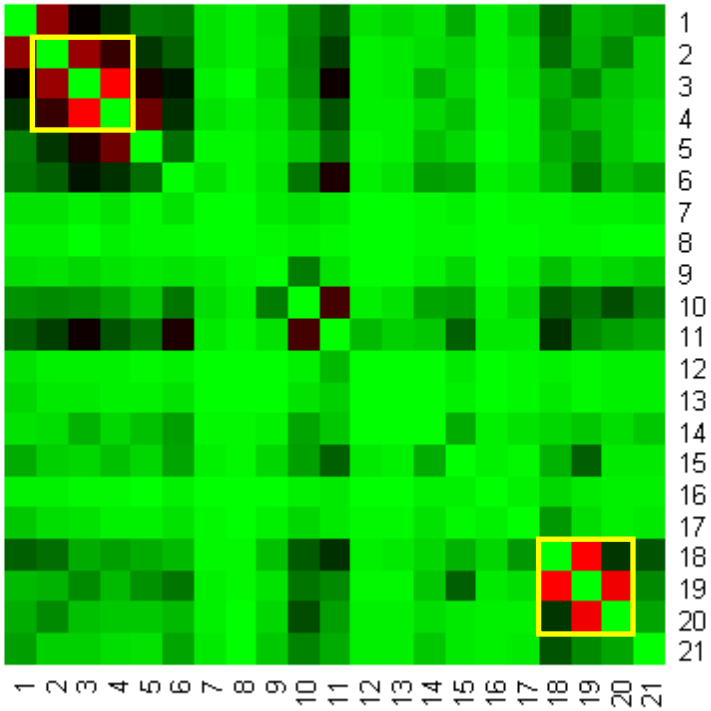


Figure S3. Illustration of the unbiased exhaustive survey of all pairs of positions within the motif. Larger differences (in darker color) indicate higher dependency. Dependent positions are illustrated in the box on the logo plot and the heatmap. The logo plots are generated using R package "seqLogo" (3). The subgraphs of multi-nucleotide logo plots were generated using a program that we modified from SeqLogo (please see Section 5 in the Supplementary Material for more details). To make the logo plots more readable, we changed the range for y-axis from 0 - 2 to 0 - 3 in the subfigures for multi-nucleotide logo plot.

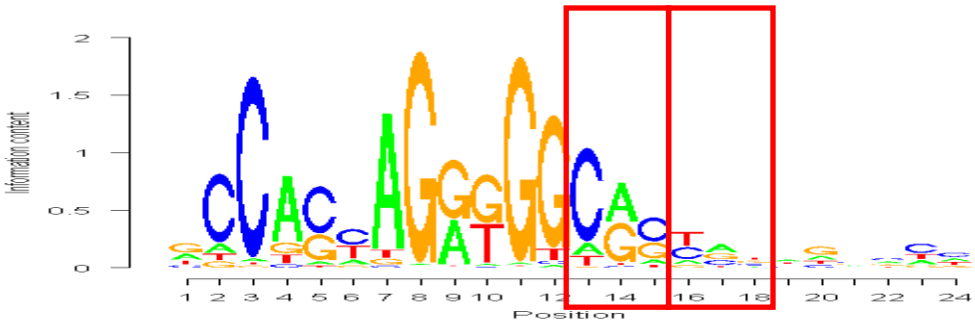
NRSF: (2,3,4) (18,19,20)

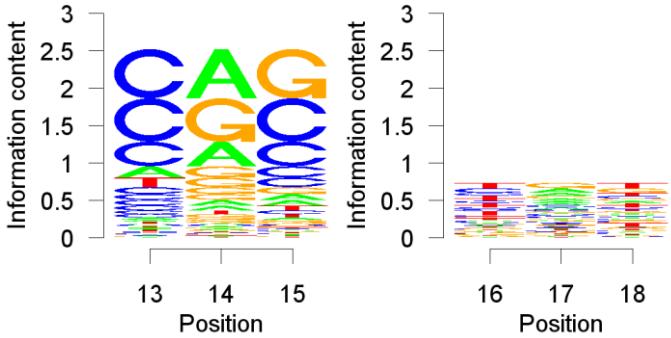


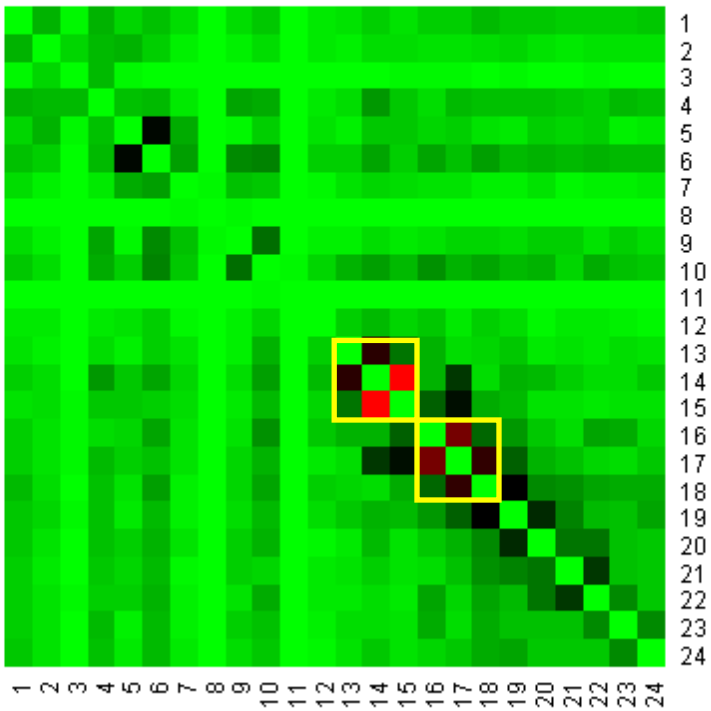




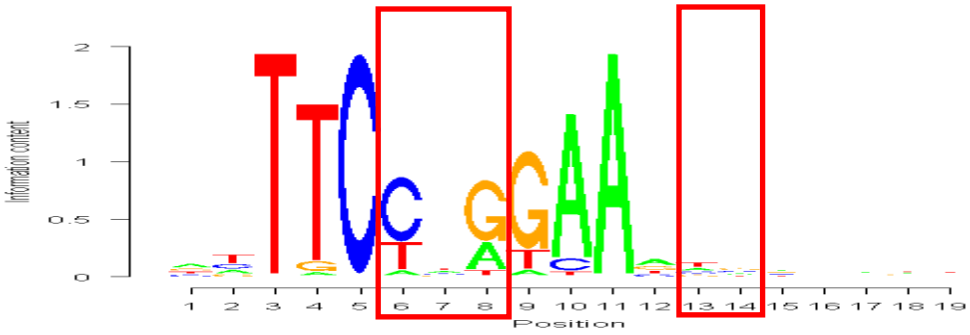
CTCF: (13,14,15) (16,17,18)

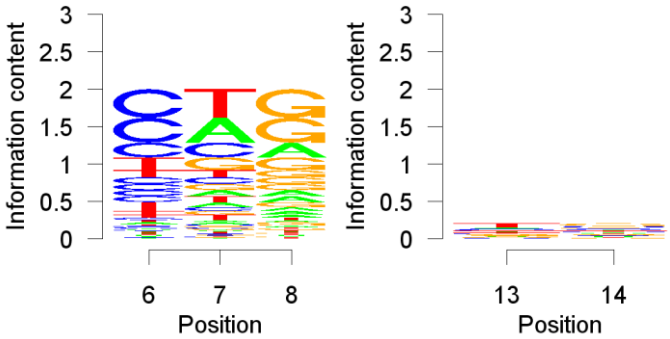






STAT1: (6,7,8) (13,14)





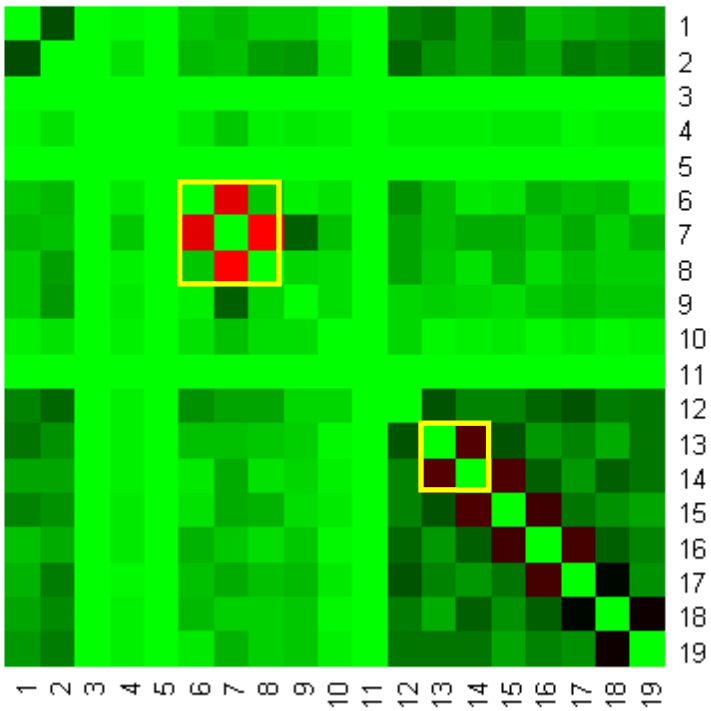


Figure S4. Difference between the observed dinucleotide frequencies and the expected dinucleotide frequencies assuming that the two positions are independent. The height of each bar is the accumulative difference in 16 dependent position pairs we identified from the four motifs.

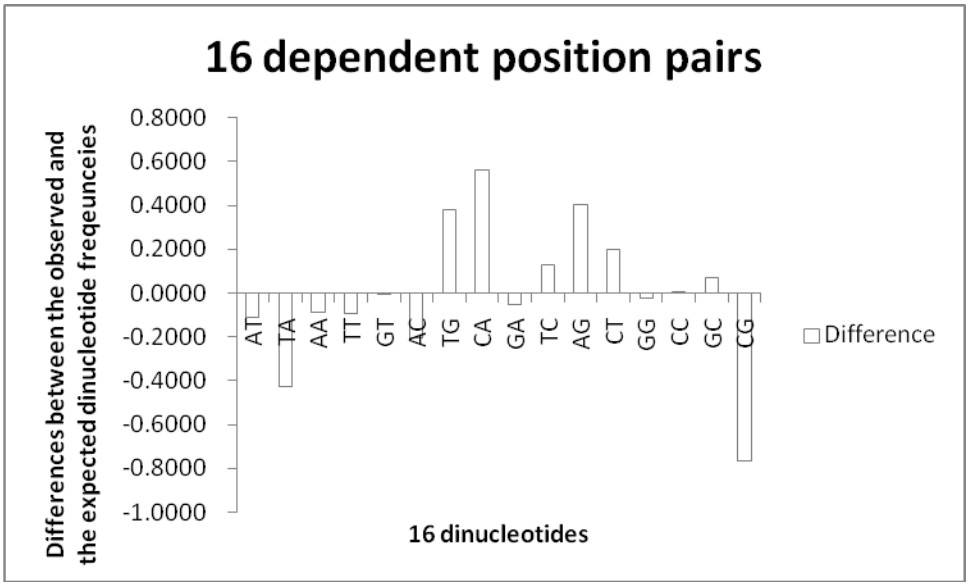
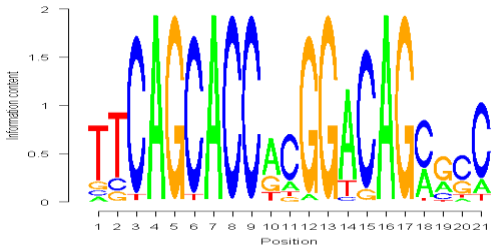


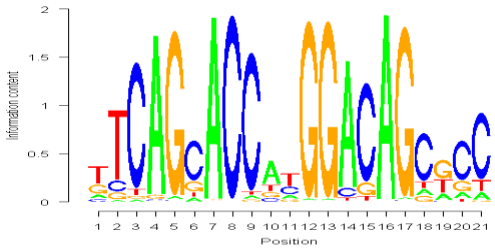
Figure S5. Comparison of NRSF motif patterns identified by different *de novo* motif finding tools as well as the known NRSF motif found in the MatBase (Genomatix Software GmBH, Munich, Germany). A. Logo plots of motifs identified by various motif finding programs as well as the NRSF motif stored in the MatBase. **B.** Comparison of motif enrichment in ChIP-Seq for five different motif finding strategies using cross validation. Training sets, testing sets and control sets were generated following the scheme described in the Method section (see "Performance Evaluation Using Real Data" in the manuscript). **C.** Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq.

A.

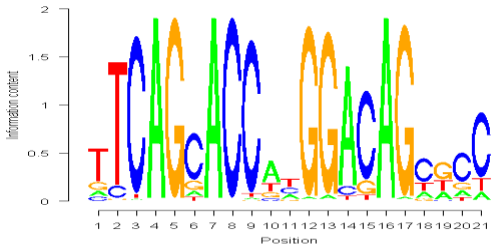
Genomatix:



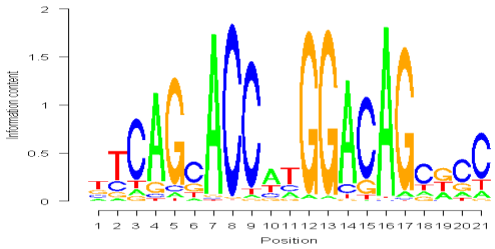
MEME:



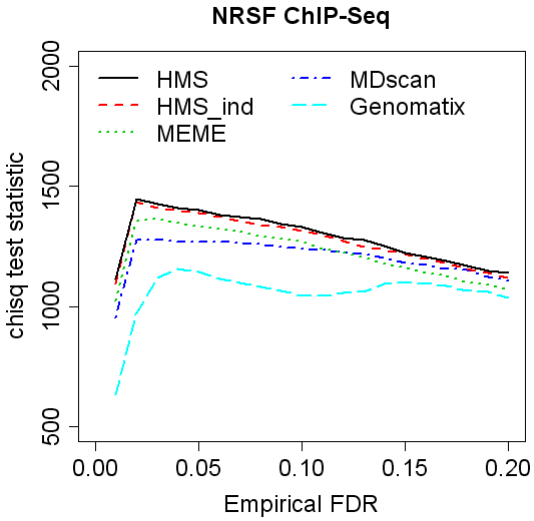
MDscan:



HMS:



B.



C.

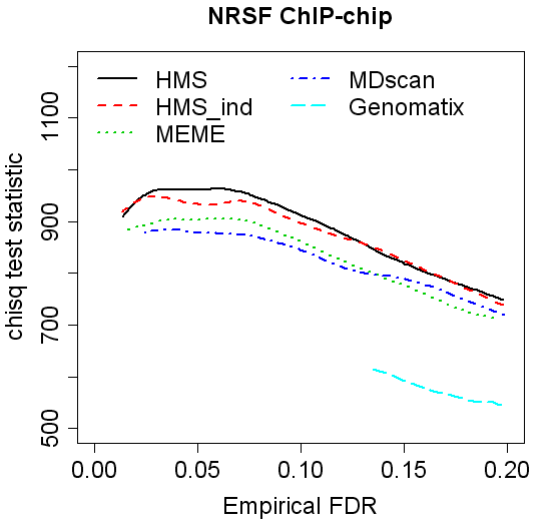
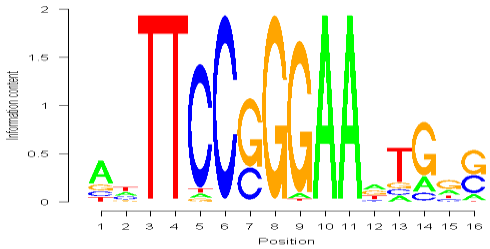


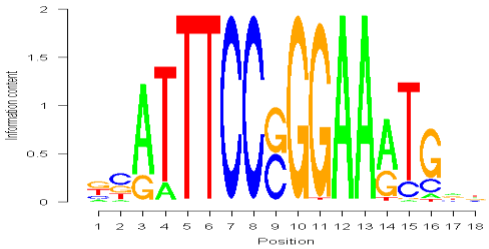
Figure S6. Comparison of STAT1 motif patterns identified by different *de novo* motif finding tools as well as known STAT motif patterns stored in the MatBase (Genomatix Software GmBH, Munich, Germany). A. Logo plots of motifs identified by various motif finding programs as well as the STAT motifs stored in the MatBase. **B.** Comparison of motif enrichment in ChIP-Seq for five different motif finding strategies using cross validation. Training sets, testing sets and control sets were generated following the scheme described in the Method section (see "Performance Evaluation Using Real Data" in the manuscript). **C.** Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq. Note: the x axis in STAT1 ChIP-chip figure is from 0 to 1.0 instead of the usual range of 0 to 0.2 due to its high empirical FDR.

A.

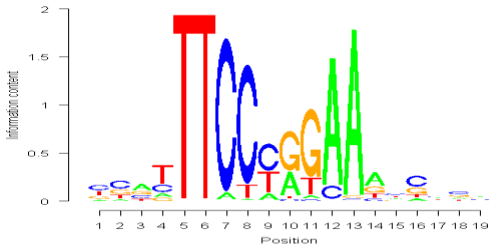
Genomatix V$STAT01:



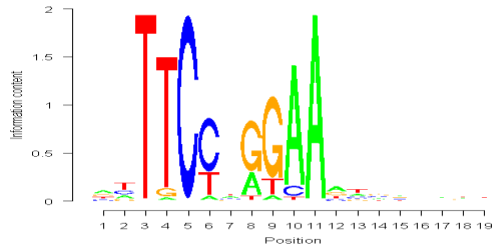
Genomatix V$STAT03:



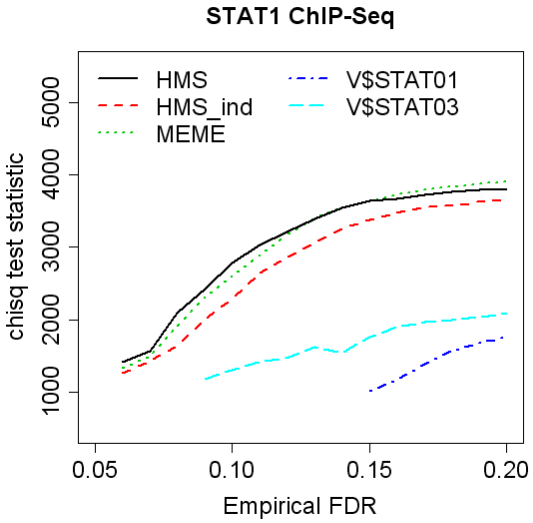
MEME:



HMS:



B.



C.

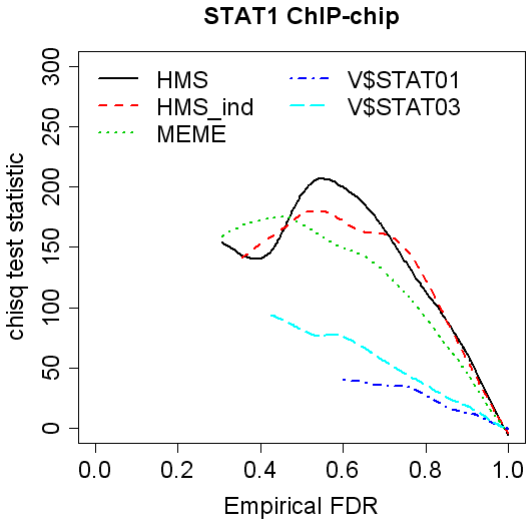
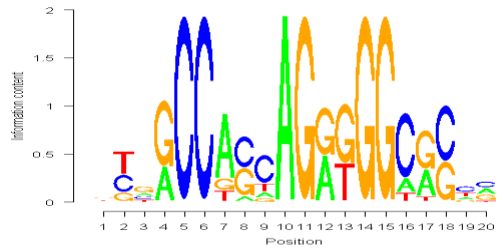


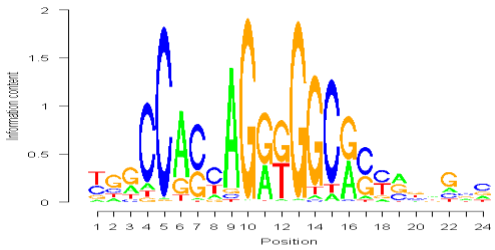
Figure S7. Comparison of CTCF motif patterns identified by different *de novo* motif finding tools as well as a known motif pattern found in Kim et al. (2007). A. Logo plots of motifs identified by various motif finding programs as well as the one found in Kim et al. (2007). **B.** Comparison of motif enrichment in ChIP-Seq for four different motif finding strategies using cross validation. Training sets, testing sets and control sets were generated following the scheme described in the Method section (see "Performance Evaluation Using Real Data" in the manuscript). **C.** Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq.

A.

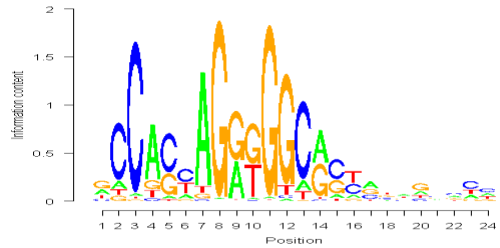
Kim07_CTCF:



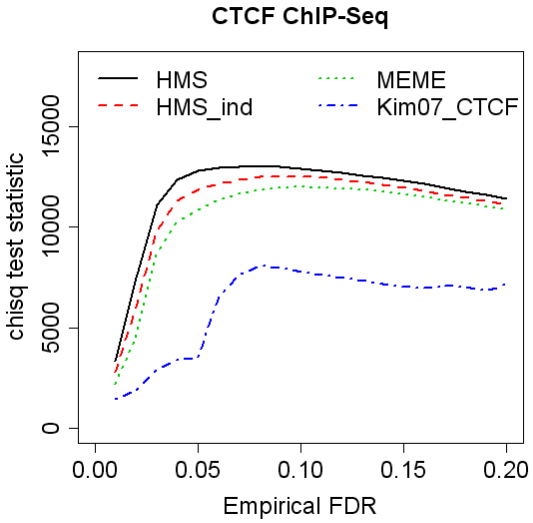
MEME:



HMS:



B.



C.

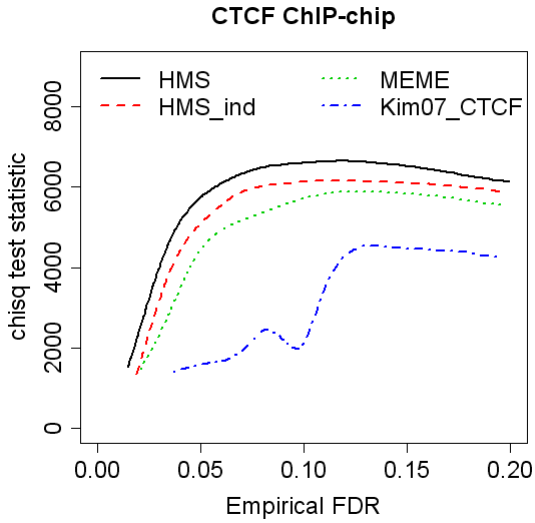
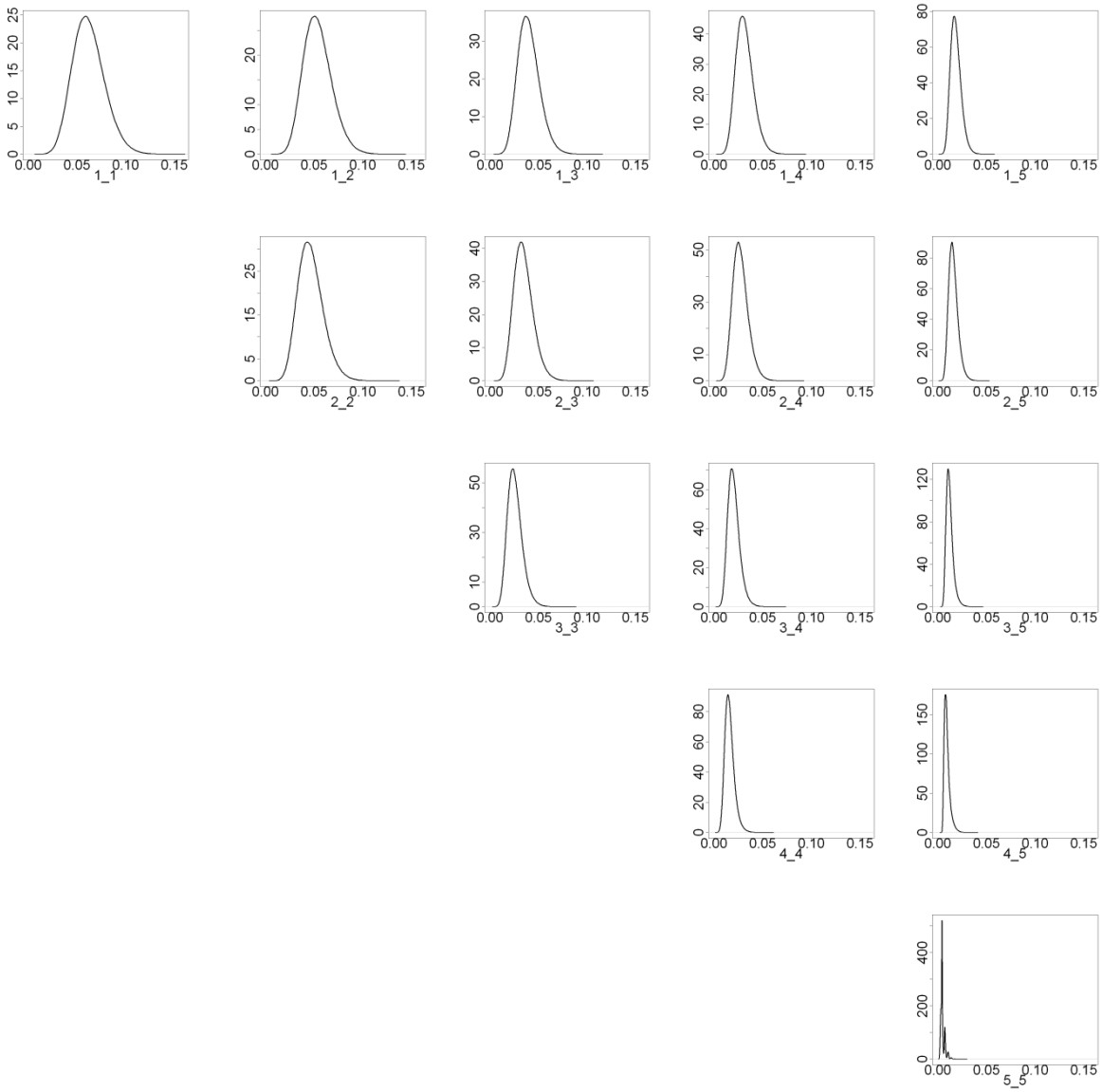


Figure S8. The histogram of 15 empirical distributions of Hamming distance. The label *i_j*, $1 \le i \le j \le 5$ shown under each of the 15 empirical distributions of Hamming distance indicates the 15 combinations of information content. Small *i*, *j* represent low information content, while large *i*, *j* represent high information content.



References

1. **Consul, P.C. (1989) *Generalized Poisson Distributions*. Marcel Dekker, New York.**
2. **Johnson, N.L., Kotz, S. and Kemp, A.W. (1992) *Univariate discrete distributions*. 2nd ed. John Wiley & Sons, New York.**
3. **Bembom, O. (2007) Sequence logos for DNA sequence alignments.**
4. **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. *Nat Biotechnol*, 20, 835-839.**
5. **Zhou, Q. and Liu, J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20, 909-916.**