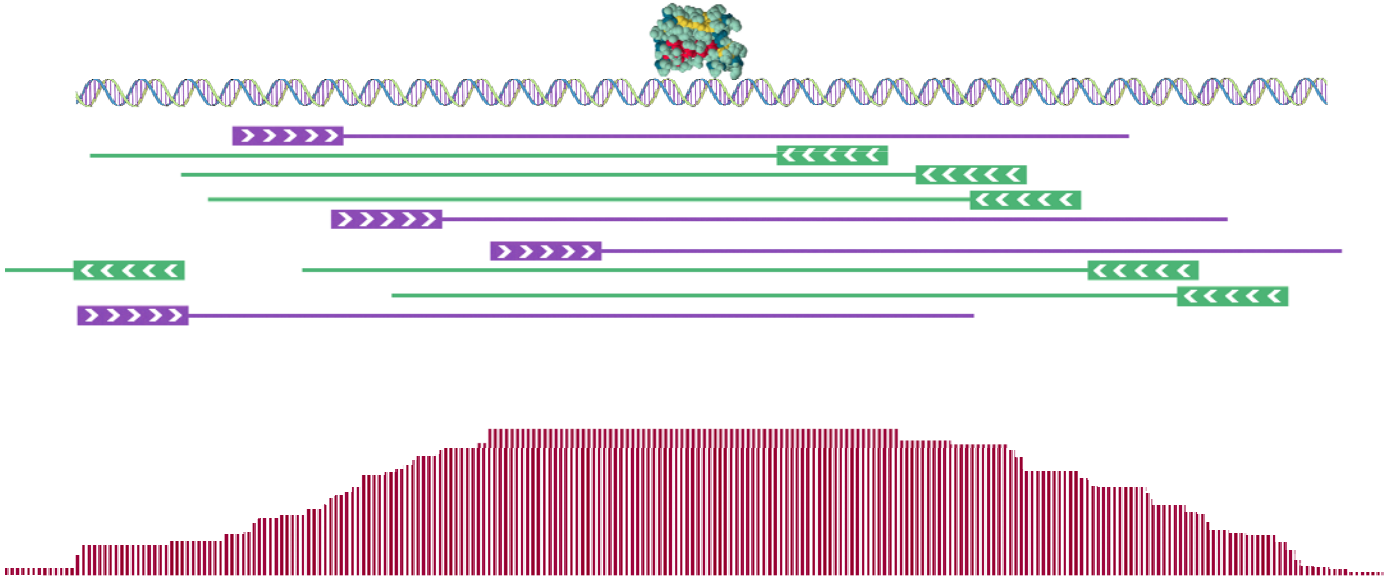
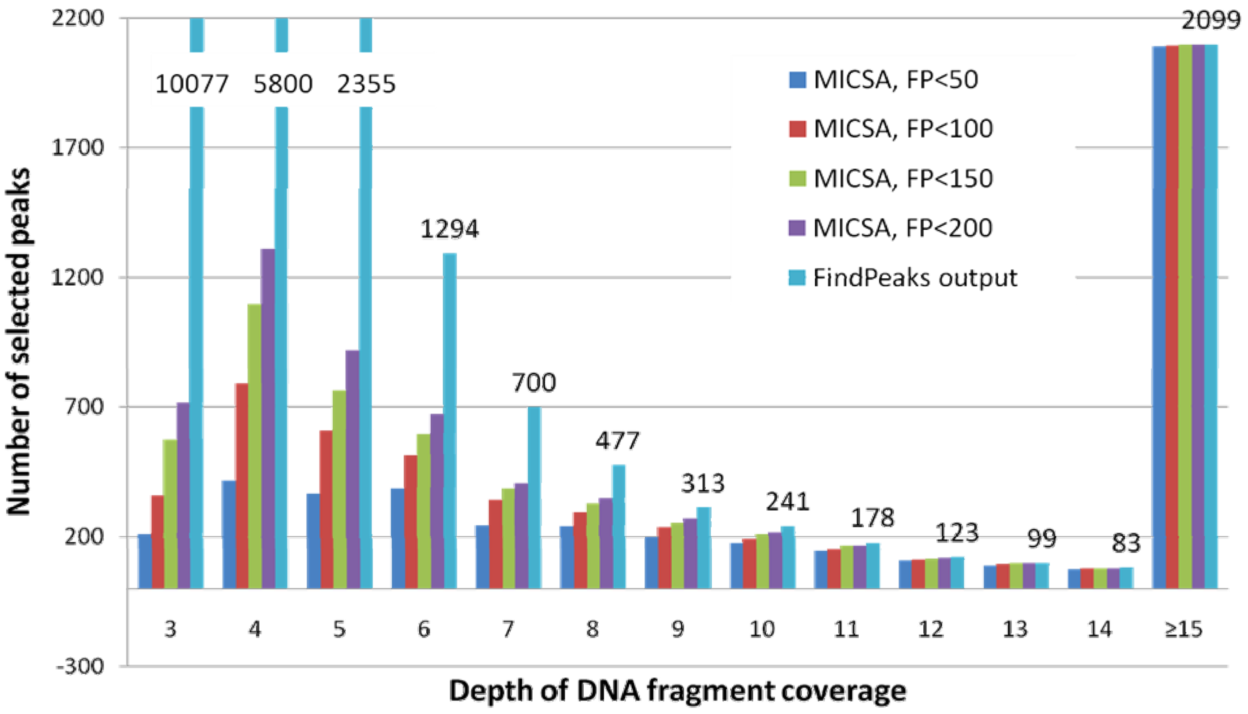
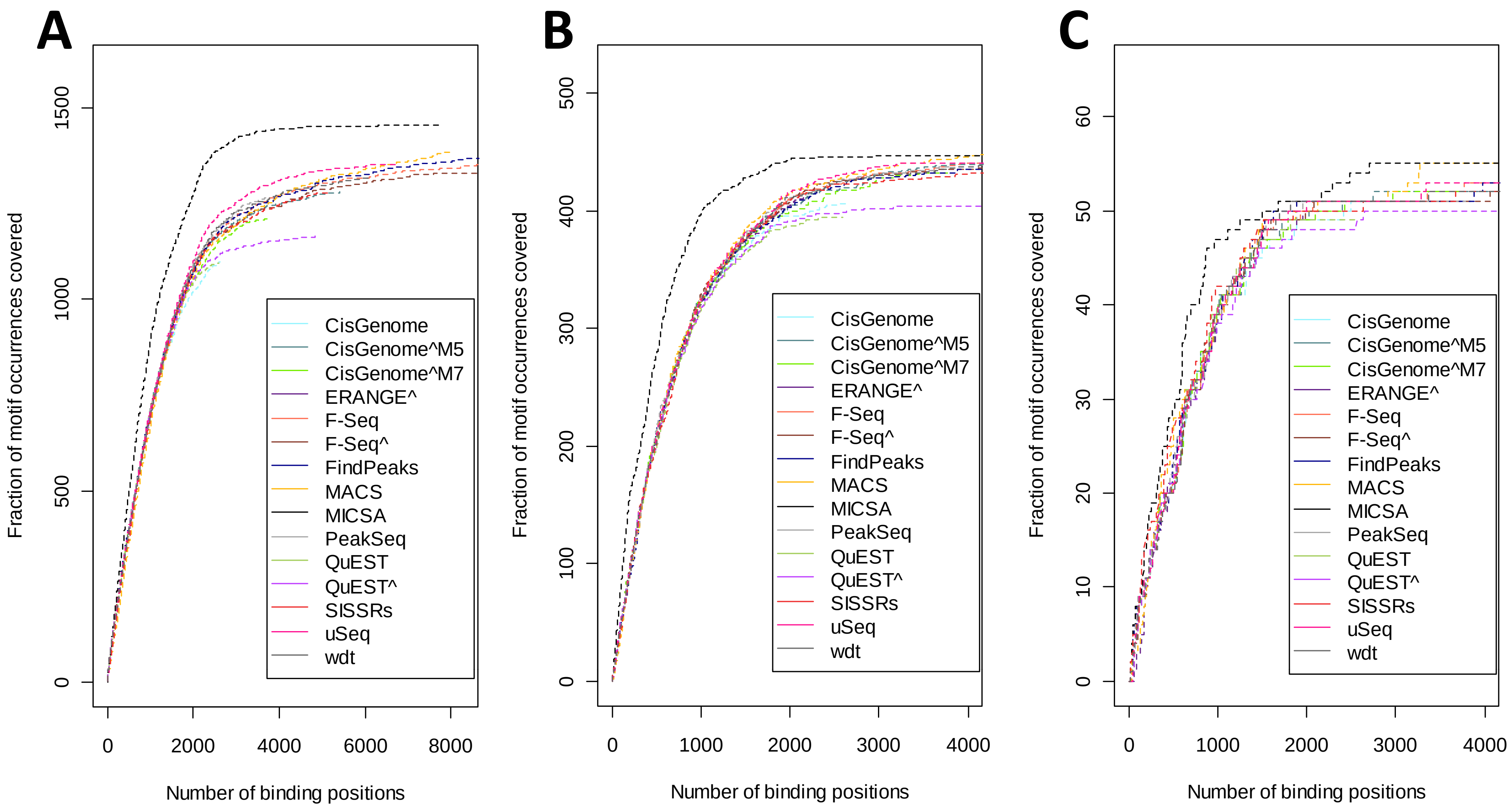


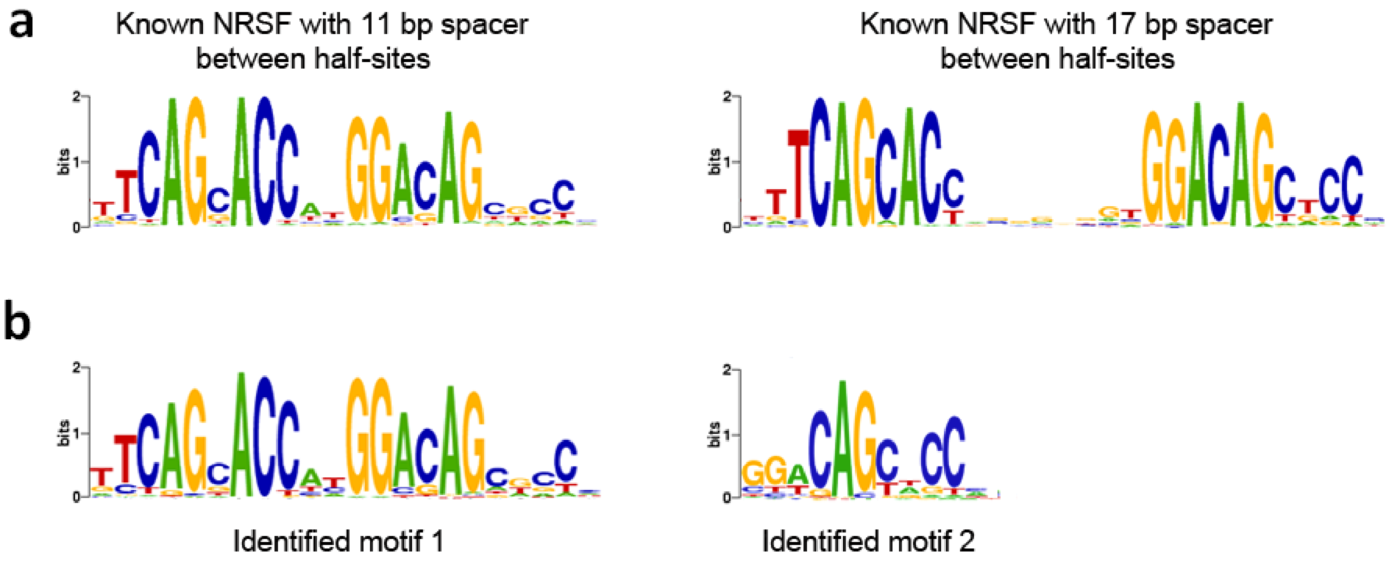
Supplementary Figure 1. Graphical representation of mapped reads for a transcription factor binding site. Each aligned read is represented by a green of violet rectangle. White arrows show the direction in which the read has been sequenced from DNA fragment. Each tag is extended to the approximate length of DNA fragment (thin green and violet lines). The density profile (red) reflects the number of overlapping DNA fragments covering a given position. The top area of the density profile usually corresponds to a protein binding site.



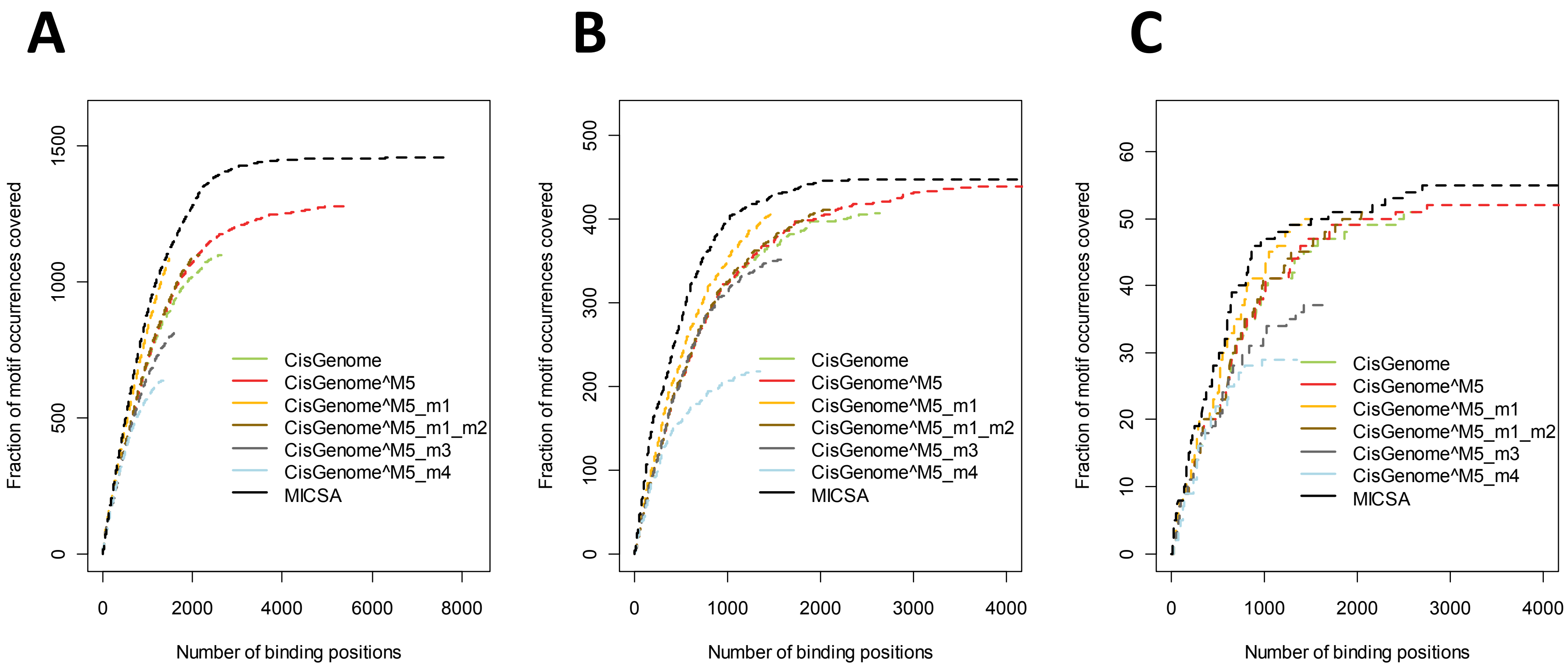
Supplementary Figure 2. Counts of peaks reported by FindPeaks (default parameters) and by MICSA (different values of maximal number of false positives *FP*) for the NRSF dataset (Johnson et al., 2007). Peak of high depth of coverage (≥15) are almost totally kept by MICSA with variable cut-off values on the number of false positives. The important difference between MICSA's and FindPeaks' output is observed for low peaks.



Supplementary Figure 3. Performance comparison of MICSA with 10 published algorithms. As a positive set of binding sites of NRSF we used (**a**) 3,000 best matches of the canonical NRSF matrix in the human genome, (**b**) 500 best matches of the canonical NRSF matrix in the human genome, (**c**) 83 q-PCR verified NRSF binding sites in the human genome. Peaks extracted by each algorithm were ranked according to in-built scores or p-values. For each number of top peaks the frequency of identified positive sites among them was plotted. “ToolName^” means that the default parameters of the tool were modified to make it report more peaks.



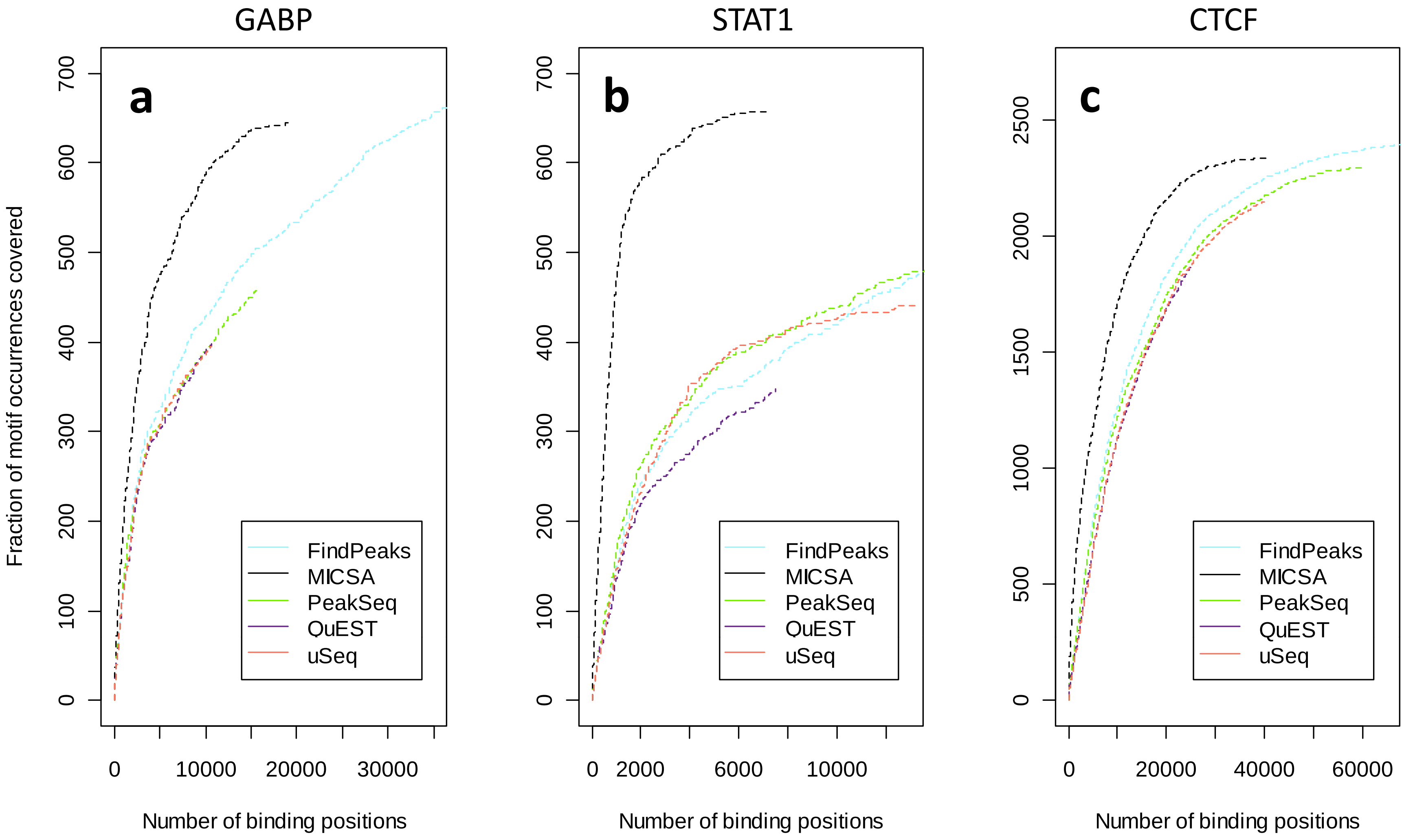
Supplementary figure 4. Motifs for neuron-restrictive silencer factor (NRSF) binding sites. (**a**) Known motifs NRSF with 11 bp and 17 bp spacer between half-sites (Johnson *et al*., 2007); (**b**) motifs identified by MICSA in the NRSF ChIP-Seq dataset. The strongest identified motif (left panel) corresponds to the first 22 positions of the known NRSF binding motif. The second identified motif matches at once the left and the right half-sites of the known NRSF motif.



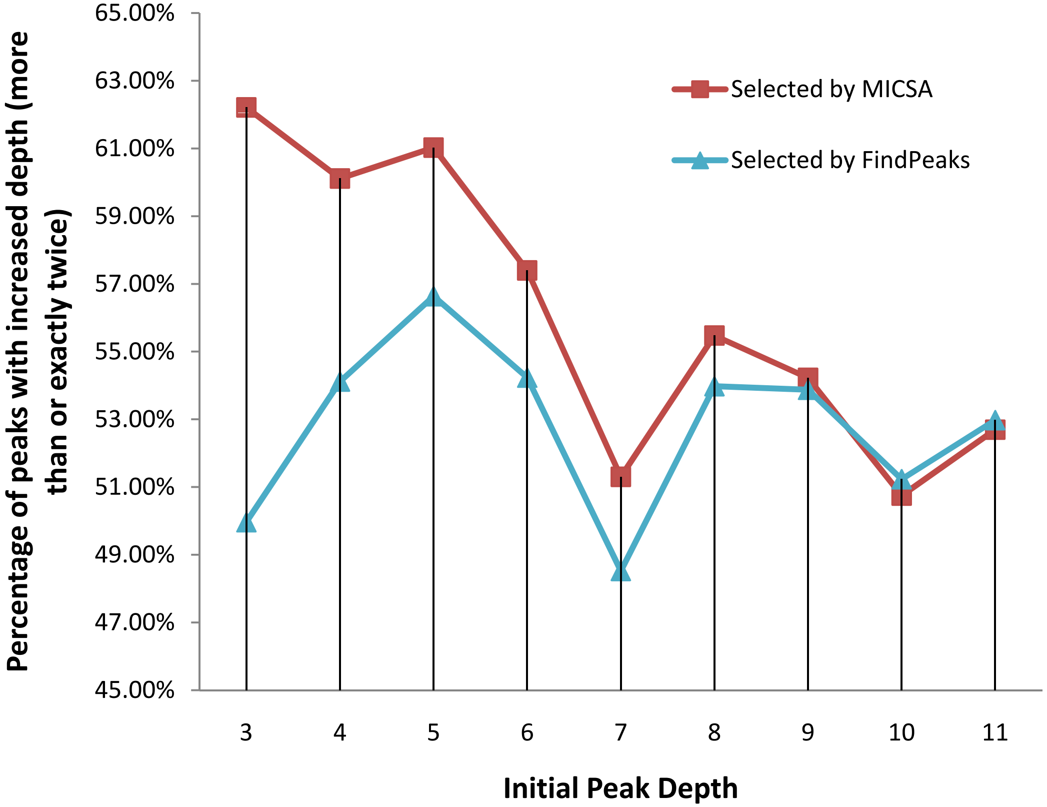
Supplementary Figure 5. Performance comparison of MICSA with motif finding module of Cis-Genome. To identify motifs in the CisGenome output we selected a set of 300 peaks with the highest max|FC| value. Running flexmodule_motif of Cis-Genome on them resulted in identification of two non-trivial motifs (consensus for **m1**: TTCAGCACCATGGACAG, Motif Score 5.3; consensus for **m2**: CCCTGGTGCTGAA, Motif Score 3.4;). In the same subset of sequences but with repetitive regions masked by Repeat Masker, we identified only one motif with a similarity to the known binding motif of NRSF (consensus for **m3**: ATGGACAGCGCC, Motif Score 4.7). If we run flexmodule_motif on the whole set of peaks, all identified motifs are low-complexity. If we mask repeats in the whole set of peaks and then run flexmodule_motif, then only one non-trivial motif is identified (consensus for **m4**: TTTCTGTGCCAT, Motif Score 3.8).

"CisGenome": CisGenome was used with its default parameters without motif filtering.
"CisGenome^M5": CisGenome was used with '-m 5' option to make it report more peaks. No motif filtering.
"CisGenome^M5_m1": CisGenome was used with '-m 5' option. Dataset containing only peaks with motifs m1 .
"CisGenome^M5_m1_m2": CisGenome was used with '-m 5' option. Dataset containing only peaks with motifs m1 or m2.
"CisGenome^M5_m3": CisGenome was used with '-m 5' option. Dataset containing only peaks with motifs m3 .
"CisGenome^M5_m4": CisGenome was used with '-m 5' option. Dataset containing only peaks with motifs m4 .
"MICSA": peaks selected by MICSA.

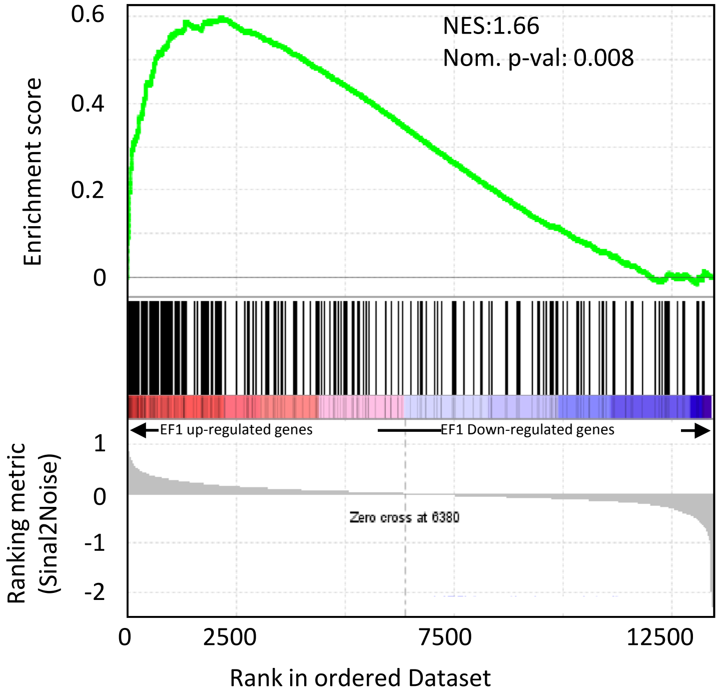
As a positive set of binding sites of NRSF we used (**a**) 3,000 best matches of the canonical NRSF matrix in the human genome, (**b**) 500 best matches of the canonical NRSF matrix in the human genome, (**c**) 83 q-PCR verified NRSF binding sites in the human genome. Peaks extracted by each algorithm were ranked according to in-built scores or p-values. For each number of top peaks the frequency of identified positive sites among them was plotted.



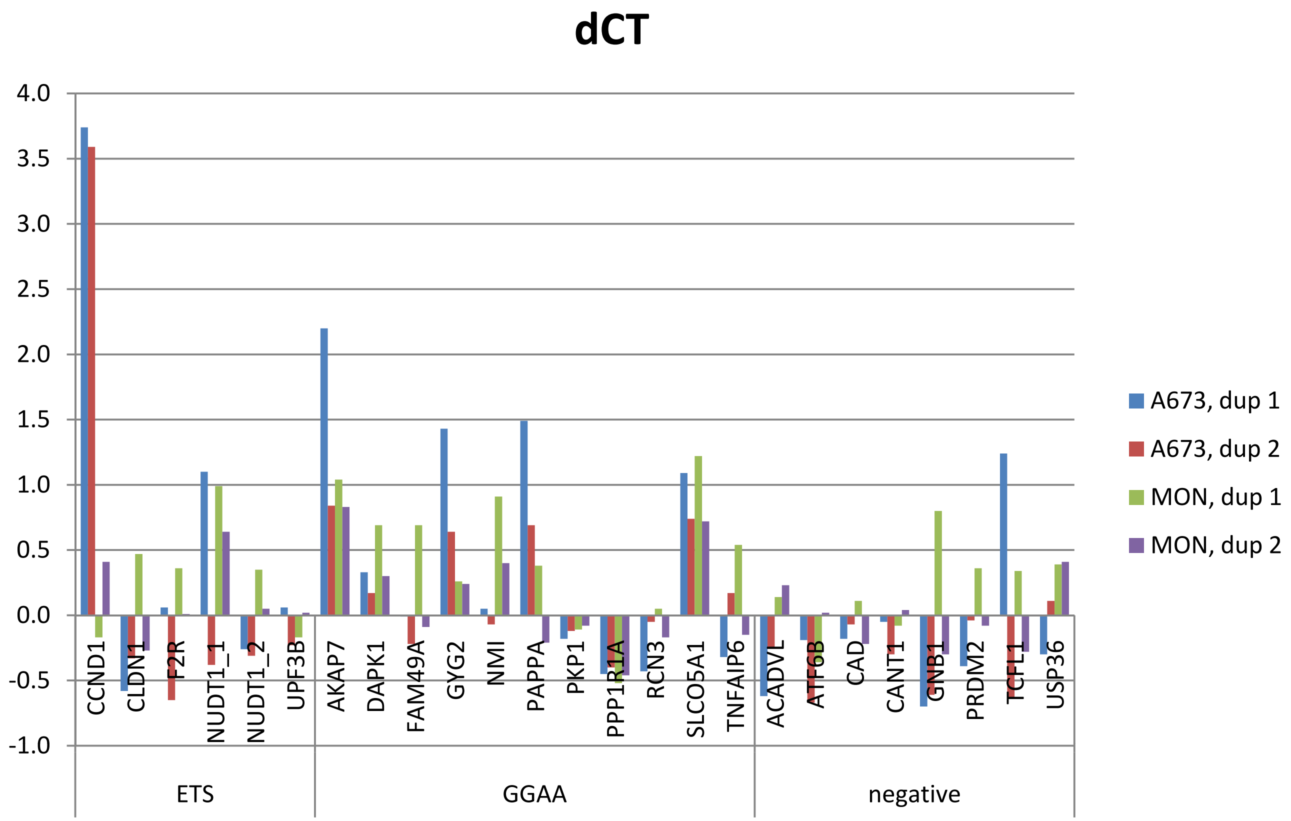
Supplementary figure 6. Results of the comparison of MICSA with FindPeaks, PeakSeq, QuEST and uSeq of three ChIP-Seq datasets: (**a**) GABP, (**b**) STAT1 and (**c**) CTCF. As a positive set of binding sites of each transcription factor we used 3,000 best matches of the corresponding canonical matrix in the human genome. Peaks extracted by each algorithm were ranked according to in-built scores or p-values. For each number of top peaks the frequency of identified positive sites among them was plotted.



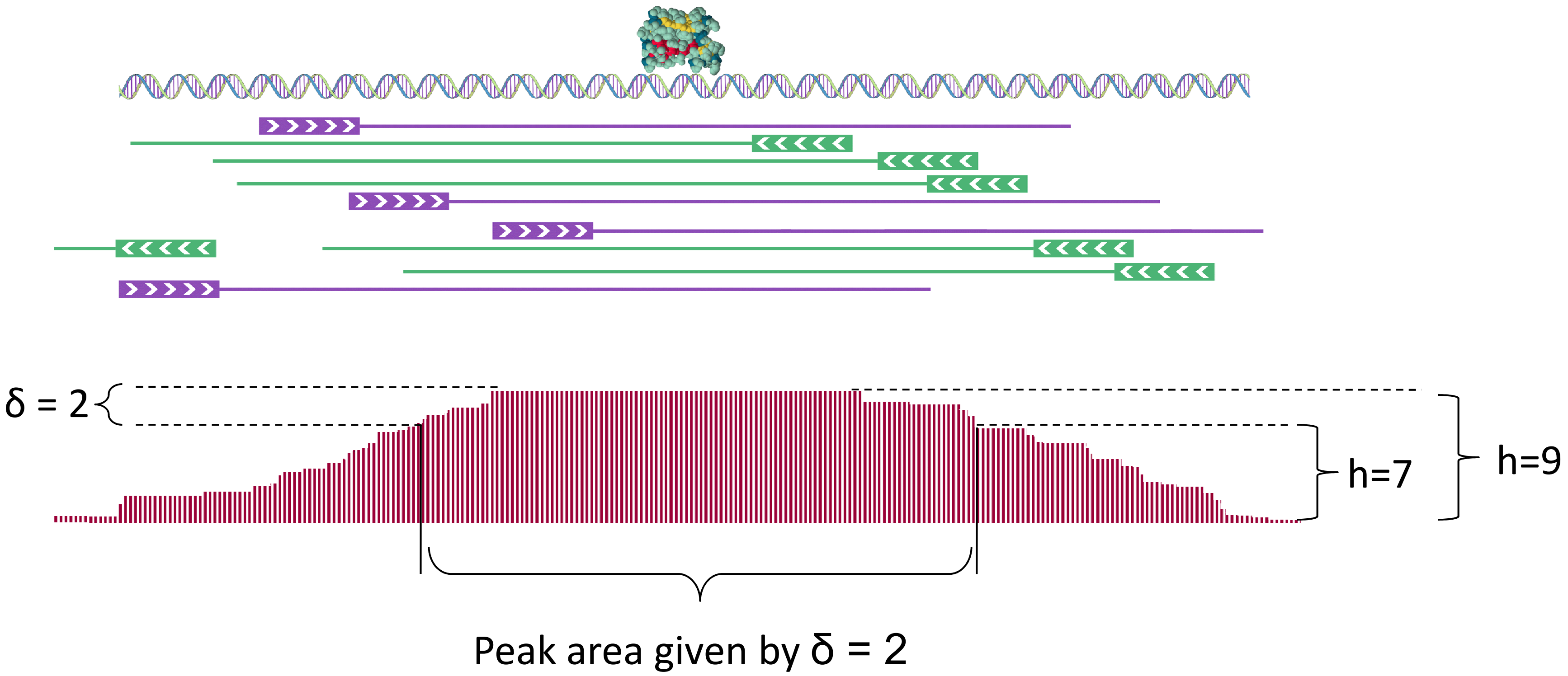
Supplementary figure 7. The percentage of peaks extracted from Jonson et al. dataset, which increase the height more than twice in the new NRSF dataset (ENCODE project, Richard M. Myers lab at the HudsonAlpha Institute for Biotechnology). Peaks selected by MICSA out of FindPeaks' output are shown in red, peaks selected by FindPeaks are shown in blue.



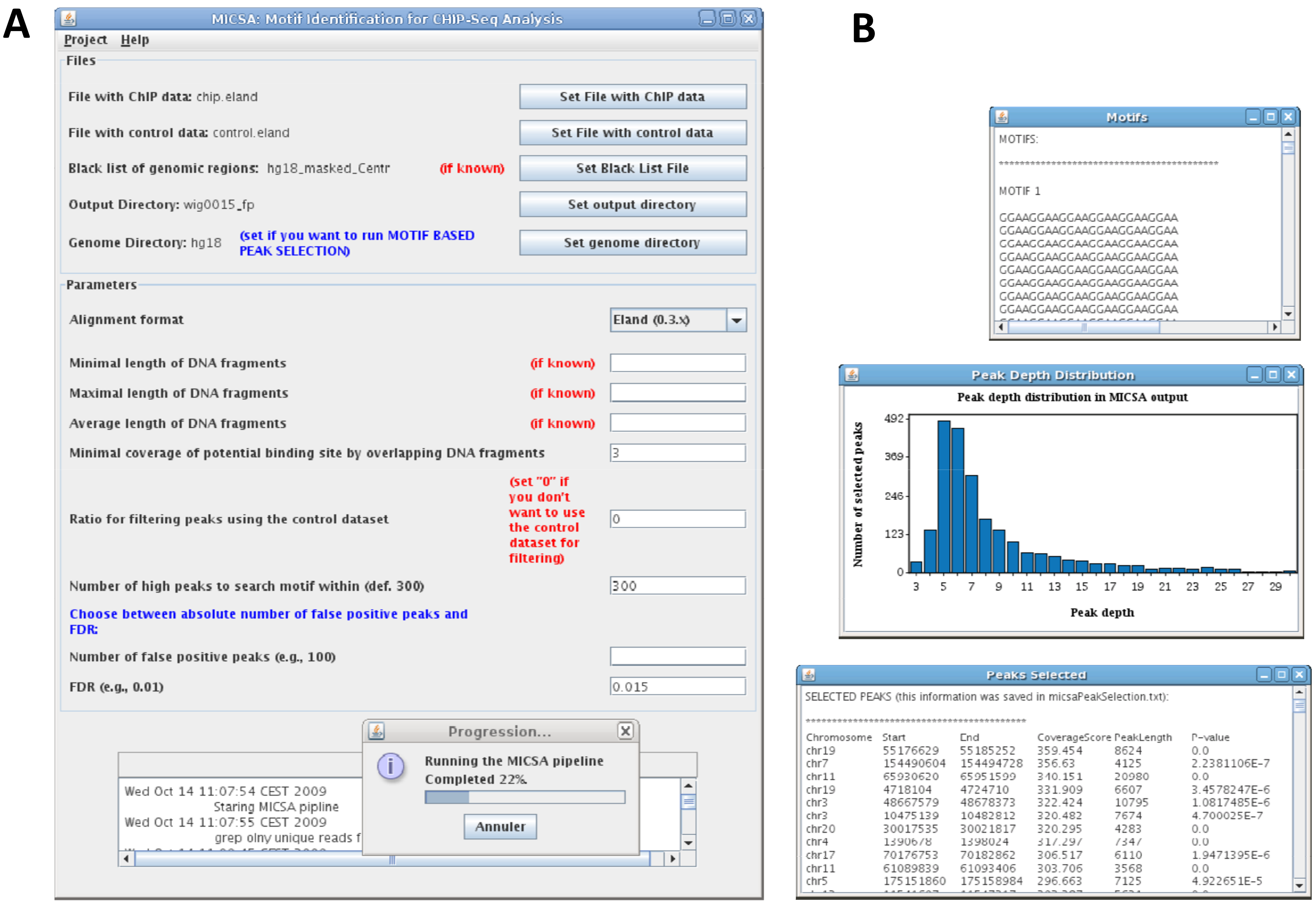
Supplementary figure 8. Gene Set Enrichment Analysis (GSEA) of genes located within 150Kb upstream or 50Kb downstream of EWS-FLI1-bound microsatellite regions. 206 identified microsatellite regions were found to have genes within -150Kb / +50Kb range, totalizing 398 genes, among which 248 genes were found to be present on HG133A Affymetrix arrays and thus used as gene set (horizontal black bars). The expression dataset resulted from previously described EWS-FLI1 inhibition experiments of A673 and SK-N-MN Ewing cell lines (Tirode *et al*., 2007; Guillon *et al*., 2009) was ranked using the signal-to-noise algorithm (grey curve at the bottom). A strong enrichment of genes flanking EWS-FLI1 bound to GGAA microsatellites among ZWS-FLI1 up-regulated genes (left side, highlighted in red) was observed. The normalized enrichment score (NES) is a ratio of actual enrichment score (ES) and the average ES for all permutations in the dataset. The nominal p-value (Nom. P-val) is a p-value of the ES of the actual gene set in a permutation test.



Supplementary figure 9. Results of ChIP-qPCR experiments validate five peaks from the MICSA's set. No peak from the control set was found to be positive in both experiments.
ETS: peaks with ETS motif; GGAA: peaks with (GGAA)n microsatelite; negative: peaks rejected by MICSA.



Supplementary Figure 10. Example of area provided by δ=2. Given this δ, the peak will be filtered if it does not contain a motif occurrence in the central area provided by δ. Here *h* is a number of overlapping DNA fragments.



File with ChIP data: chip.eland

File with control data: control.eland

Black list of genomic regions: hg18_masked_Centr (if known)

Output Directory: wig0015_fp

Genome Directory: hg18 (set if you want to run MOTIF BASED PEAK SELECTION)

Alignment format: Eland (0.3.x)

Minimal length of DNA fragments (if known)

Maximal length of DNA fragments (if known)

Average length of DNA fragments (if known)

Minimal coverage of potential binding site by overlapping DNA fragments: 3

Ratio for filtering peaks using the control dataset (set "0" if you don't want to use the control dataset for filtering): 0

Number of high peaks to search motif within (def. 300): 300

Choose between absolute number of false positive peaks and FDR:

Number of false positive peaks (e.g., 100)

FDR (e.g., 0.01): 0.015

```
Wed Oct 14 11:07:54 CEST 2009
                Staring MICSA pipline
Wed Oct 14 11:07:55 CEST 2009
                grep olny unique reads f
```

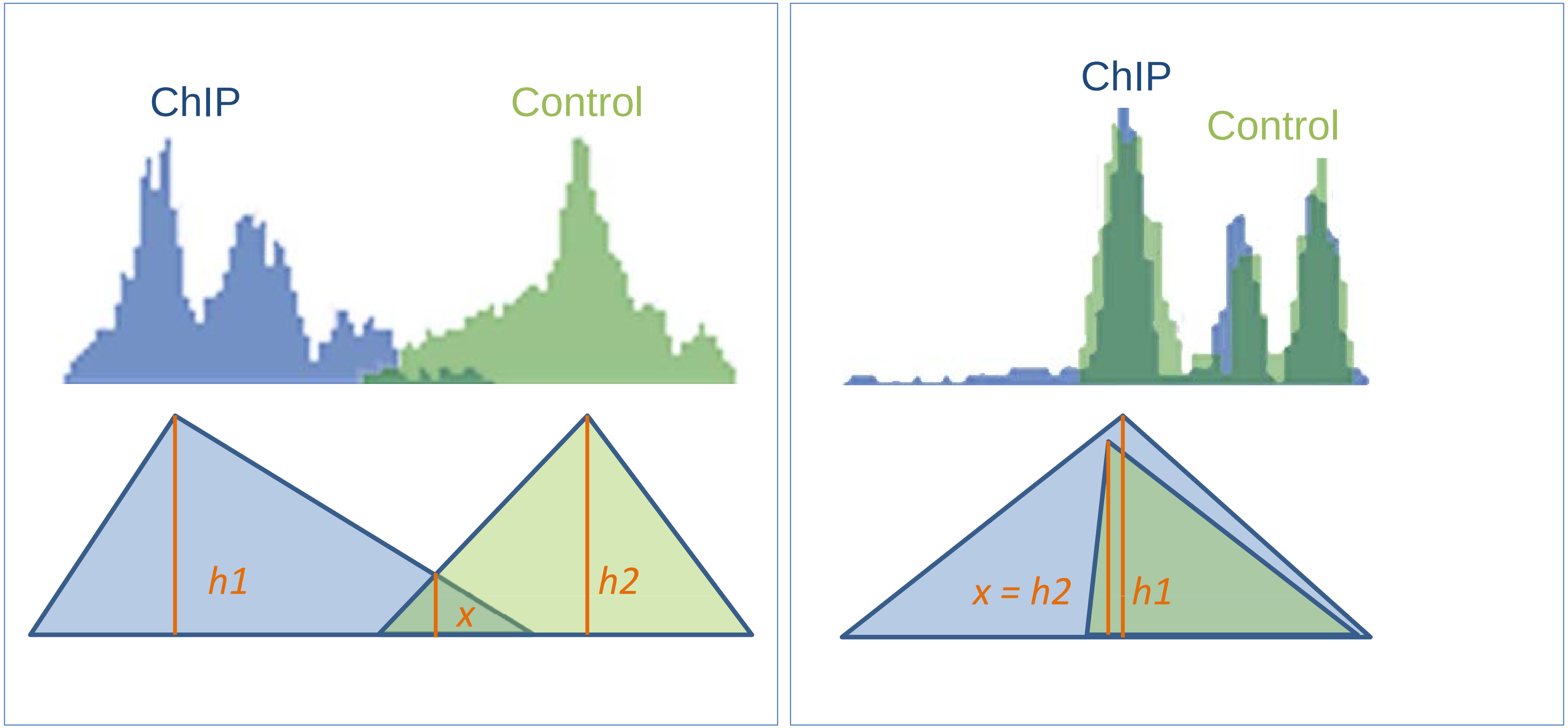
Progression...

Running the MICSA pipeline
Completed 22%.

```
MOTIFS:

********************************************

MOTIF 1

GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
```

```
SELECTED PEAKS (this information was saved in micsaPeakSelection.txt):

******************************************

Chromosome   Start       End         CoverageScore PeakLength   P-value
chr19        55176629    55185252    359.454       8624         0.0
chr7         154490604   154494728   356.63        4125         2.2381106E-7
chr11        65930620    65951599    340.151       20980        0.0
chr19        4718104     4724710     331.909       6607         3.4578247E-6
chr3         48667579    48678373    322.424       10795        1.0817485E-6
chr3         10475139    10482812    320.482       7674         4.700025E-7
chr20        30017535    30021817    320.295       4283         0.0
chr4         1390678     1398024     317.297       7347         0.0
chr17        70176753    70182862    306.517       6110         1.9471395E-6
chr11        61089839    61093406    303.706       3568         0.0
chr5         175151860   175158984   296.663       7125         4.922651E-5
```

Supplementary Figure 11. Screen shot of the MICSA graphical interface: input (A) , output of MICSA (B).



h1/*x* > threshold? ➡ Keep the peak

Supplementary Figure 12. Filtering peaks occurring both in the ChIP and control data. The actual peak shapes are replaced by triangles with corresponding start, end, maximum positions and heights. Then, the height (*x*) of maximal overlap is calculated. The ChIP peak is rejected if its height (*h1*) divided by *x* is less than or equal to the threshold defined by user (default value is 2).

Supplementary Table 1. Command lines used to run the tested software.

Tool	Command lines
CisGenome	file_eland2aln -i /…/chip1862_hg18.txt -o chip.aln file_eland2aln -i /…/mock1862_hg18.txt -o control.aln tablesorter_str chip.aln tablesorter_str control.aln hts_aln2barv2 -i chip.aln.sort -o chip.bar hts_aln2barv2 -i control.aln.sort -o control.bar hts_windowsummaryv2_2sample -i chip.bar -n control.bar -g chrlist.txt -l chrlen.txt -w 100 -o summarys2.txt hts_peakdetectorv2_2sample -i chip.bar -n control.bar -d /…/ -o outputCisGenome -f summarys2.txt.fdr -p 0.354986 hts_alnshift2bar -i chip.bar -s 152 hts_alnshift2bar -i control.bar -s 152 hts_windowsummaryv2_2sample -i chip.bar -n control.bar -g chrlist.txt -l chrlen.txt -w 100 -o summarys2.txt -z 1 hts_peakdetectorv2_2sample -i chip.bar -n control.bar -d /…/ -o outputCisGenome -f summarys2.txt.fdr -p 0.354631 -z 1
CisGenome^M5	hts_peakdetectorv2_2sample -i chip.bar -n control.bar -d /…/ -o outputCisGenome^M5 -f summarys2.txt.fdr -m 5 -p 0.354986 -z 1
CisGenome^M7	hts_peakdetectorv2_2sample -i chip.bar -n control.bar -d /…/ -o outputCisGenome^M7 -f summarys2.txt.fdr -m 7 -p 0.354986 -z 1
ERANGE	python /ERANGE3.0.2/commoncode/makerdsfrombed.py chip /…/bed/chip.txt chip.rds python /ERANGE3.0.2/commoncode/makerdsfrombed.py control /…/bed/control.txt control.rds python /ERANGE3.0.2/commoncode/findall.py er chip.rds ErangeOut -control control.rds regionoutfileER3_def -listPeak revbackground
*ERANGE^*	python /ERANGE3.0.2/commoncode/findall.py er chip.rds regminP02m3r3 -control control.rds ErangeOut^ -listPeak revbackground -minPeak 0.2 -minimum 3 -ratio 3
FindPeaks	grep 'U[012]' /…/chip1862_hg18.txt > chip.eland grep 'U[012]' /…/mock1862_hg18.txt > control.eland java –jar SeparateReads.jar eland chip.eland out_chip/ java –jar SeparateReads.jar eland control.eland out_control/ java –jar SortFiles.jar eland out_chip/ out_chip/*.part.eland.gz java –jar SortFiles.jar eland out_control/ out_control/*.part.eland.gz java -Xmx2G -jar FindPeaks.jar -aligner eland -eff_frac 0.7 -duplicatefilter -input out_chip/*.part.eland.gz -name FindPeaksOut -output /…/wig -auto_threshold -control out_control/*.part.eland.gz -dist_type 1
fseq	fseq -v -of bed -l 200 -t 12 -o /…/resT12 chip.txt cat *.bed >fseqResult.bed
*fseq^*	fseq -v -of bed -l 200 -t 12 -b /…/bff_20 -o /…/resT12b chip.txt cat *.bed >fseq^Result.bed
MACS	macs -t chip1862_hg18.txt -c mock1862_hg18.txt --name=macsNRSF --format=ELAND --pvalue=1e-5
MICSA	java -Xmx2G -jar FindPeaks.jar -aligner eland -eff_frac 0.7 -duplicatefilter -input out_control/*.part.eland.gz -name control output wig -dist_type 1 -minimum 1 java -Xmx2G -jar FindPeaks.jar -aligner eland -eff_frac 0.7 -duplicatefilter -input out_chip/*.part.eland.gz -name chip output /wig -dist_type 1 -minimum 3 java -Xmx2G DeleteRegions -f wig/chip_triangle_standard.peaks -r hg18_masked_Centr; java -Xmx2G DeleteRegions -f wig/control_triangle_standard.peaks -r hg18_masked_Centr java -Xmx2G Summary -f wig/chip_triangle_standard.peaks -c wig/control_triangle_standard.peaks -r 0.67 java -Xmx3G -jar micsa.jar -name micsaResults.txt -f wig/chip_triangle_standard.peaks -fdr 0.05 -o /results/ -l FindPeaksSummary.txt –g /HumanGenome/ -w wig/chip_triangle_standard.wig.gz

PeakSeq	No command line; perl and sh file need to be changed manually. Parameter:
	my $pval_threshold = 0.05;
	my $L = 200;my $bin_size = 10000; my $bin_sizeM = 1000; my $max_count = 3; my $extended_region_size = 2000; my $Pf = 1;my $eland_filename = "sample/chip1862_hg18_eland_result.";
	my $input_filename = "input/mock1862_hg18_eland_result.« ;my $bed_filename = "bed_files/Pchip1862_hg18_sites.« ;
	my $sgr_filename = "sample/chip1862_hg18."; my $L = 200;
	my $window_size = 1000000; my $max_threshold = 100; my $max_gap = 200; my $FDR_required = 0.05; my $number_of_sims = 10;
	my $map_filename = "Mapability_HG.txt“;

QuEST	Perl generate_QuEST_parameters.pl -eland_align_ChIP /…/chip1862_hg18.txt -eland_align_RX_noIP /…/mock1862_hg18.txt -gt genome_table –ap /…/results -ChIP_name questResults –advanced perl run_QuEST_with_param_file.pl -ap /…/results
*QuEST^*	perl generate_QuEST_parameters.pl -eland_align_ChIP /…/chip1862_hg18.txt -eland_align_RX_noIP /…/mock1862_hg18.txt -gt genome_table -ap /…/results -ChIP_name quest^Results -advanced perl run_QuEST_with_param_file.pl -ap /…/results
sissrs	
uSeq	java -Xmx1500M –jar /USeq/Apps/Tag2Point -f /…/BedFiles/ -v H_sapiens_Mar_2006 java -Xmx1500M -jar /USeq/Apps/FilterPointData -p /…/BedFiles/control_Point -r /…/hg18_masked_Sat.bed -a 0.75 -s 50 java -Xmx1500M -jar /USeq/Apps/FilterPointData -p /…/BedFiles/chip_Point -r /…/hg18_masked_Sat.bed -a 0.75 -s 50 java -Xmx1500M –jar /USeq/Apps/PeakShiftFinder -t /…/BedFiles/chip_Point_hg18_masked_Sat_Filt50bp/ -c /…/BedFiles/control_Point_hg18_masked_Sat_Filt50bp/ -s /…/Results/ -a 5 java -Xmx1500M -jar /USeq/Apps/ScanSeqs -t /…/BedFiles/chip_Point_hg18_masked_Sat_Filt50bp/ -c /…/BedFiles/control_Point_hg18_masked_Sat_Filt50bp/ -s /…/Results/ -w 200 -p 0 java -Xmx500M -jar /USeq/Apps/EnrichedRegionMaker -f /…/Results/ -i 1,2,4 -s 20,13,1 -t /…/BedFiles/chip_Point_hg18_masked_Sat_Filt50bp/ -c /…/BedFiles/control_Point_hg18_masked_Sat_Filt50bp/

wdt	R command line:
	library(spp);library(snow);cluster <- makeCluster(4);
	chip.data <- read.eland.tags("chip1862_hg18.txt",max.eland.tag.length=25); #2290691
	input.data <- read.eland.tags("mock1862_hg18.txt",max.eland.tag.length=25); #2370500
	chip.data <- select.informative.tags(chip.data,binding.characteristics);
	input.data <- select.informative.tags(input.data,binding.characteristics);
	chip.data <- remove.local.tag.anomalies(chip.data);
	input.data <- remove.local.tag.anomalies(input.data);
	tag.shift <- round(binding.characteristics$peak$x/2);fdr <- 1e-2;
	detection.window.halfsize <- binding.characteristics$whs;
	bp <- find.binding.positions(signal.data=chip.data,control.data=input.data,fdr=fdr,whs=detection.window.halfsize,cluster=cluster)
	output.binding.results(bp,"example.binding.positions.txt")

Supplementary Table 2. Performance comparison of MICSA with 10 published algorithms. 3000 best matches of the canonical NRSF matrix in the human genome are used as positive set of binding sites of NRSF. Only top 3000 peaks predicted by each tool are considered. "*" means that with these parameters the tool reported less than 3000 peaks. "ToolName^" means that the default parameters of tool were modified to make it report more peaks.

Tool	Absolute number of identified sites	Percentage
MICSA	1422	47.40%
uSeq	1254	41.80%
wdt	1229	40.97%
PeakSeq	1227	40.90% *
F-Seq	1217	40.57%
FindPeaks	1216	40.53%
CisGenome^M5	1203	40.10%
*F-Seq^*	1199	39.97%
MACS	1195	39.83%
SISSRs	1194	39.80%
CisGenome^M7	1184	39.47% *
*QuEST^*	1132	37.73%
*ERANGE^*	1118	37.27%
CisGenome	1098	36.60% *
QuEST	1096	36.53% *
ERANGE	1063	35.43% *

Supplementary Table 3. Performance comparison of MICSA with 10 published algorithms. 500 best matches of the canonical NRSF matrix in the human genome are used as positive set of binding sites of NRSF. Only top 3000 peaks predicted by each tool are considered. '*' means that with these parameters the tool reported less than 3000 peaks. "ToolName^" means that the default parameters of tool were modified to make it report more peaks.

Tool	Absolute number of identified sites	Percentage
MICSA	446	89.20%
uSeq	437	86.19%
MACS	435	85.80%
F-Seq	432	85.21%
CisGenome^M5	431	85.01%
*F-Seq^*	431	85.01%
PeakSeq	431	85.01% *
wdt	430	84.81%
CisGenome^M7	428	84.42% *
FindPeaks	428	84.42%
SISSRs	424	83.63%
*ERANGE^*	412	81.26%
CisGenome	406	81.20% *
ERANGE	403	80.60% *
*QuEST^*	402	79.29%
QuEST	394	77.71% *

Supplementary Table 4. Performance comparison of MICSA with 10 published algorithms. 83 qPCR verified NRSF binding sites are used as positive set of binding sites of NRSF. Only top 3000 peaks predicted by each tool are considered. ‘*’ means that with these parameters the tool reported less than 3000 peaks. “ToolName^” means that the default parameters of tool were modified to make it report more peaks.

Tool	**Absolute number of identified sites**	**Percentage**
MICSA	55	66.27%
CisGenome^M7	52	62.65% *
CisGenome^M5	52	62.65%
MACS	52	62.65%
CisGenome	51	61.45% *
FindPeaks	51	61.45%
SISSRs	51	61.45%
F-Seq	51	61.45%
*F-Seq^*	51	61.45%
uSeq	51	61.45%
wdt	51	61.45%
PeakSeq	51	61.45% *
*SISSRs^*	51	61.45%
*QuEST^*	50	60.24%
*ERANGE^*	50	60.24%
ERANGE	50	60.24% *

Supplementary Table 5. Comaprison of peak depth in the datasets published by Johnson et al.,2007, and the dataset from R.M. Myers lab (ENCODE project). Peaks selected by MICSA and by FindPeaks.

	Number of peaks with depth increased MORE than twice in the dataset from Richard M. Myers lab		**Number of peaks with depth increased LESS than twice in the dataset from Richard M. Myers lab**	
Peak depth in Johnson et al. dataset	**MICSA output**	**FindPeaks output**	**MICSA output**	**FindPeaks output**
3	540	4521	328	4528
4	755	2873	501	2436
5	584	1314	373	1006
6	454	692	337	584
7	256	328	243	348
8	228	251	183	214
9	154	160	130	137
10	102	104	99	99
11	88	89	79	79
12	67	69	43	43
13	64	67	33	33
14	46	46	27	28
15	48	48	36	36
16	50	50	17	17
17	38	38	18	18
18	36	36	13	13
19	44	44	7	7
20	40	40	16	16
21	30	30	12	12
22	29	29	12	12

Supplementary Table 6. Possible direct targets of EWS-FLI1, genes which have a predicted EWS-Fli1 binding sites near transcription start site and are regulated by EWS-Fli1. The whole list of genes is available on the MICSA web-site.

Gene	Distance to TSS	Fold change	Peak coordinates	Site type
ABHD6	-6344	8.633878908	chr3:58191801-58192364	microsatellite
AKAP7	-20352	18.62834355	chr6:131487531-131487965	microsatellite
AK1	-5067	-5.42624151	chr9:129684745-129685381	ETS-site
ARHGAP1	17360	-3.290448257	chr11:46661093-46661673	ETS-site
ARHGAP1	-24071	-3.290448257	chr11:46702065-46702967	ETS-site
C16orf68	-56188	2.164065126	chr16:8566052-8567176	microsatellite
C1orf112	-232	3.524280518	chr1:168030537-168031032	ETS-site
CADPS2	-84489	2.226520652	chr7:122398123-122398517	microsatellite
CAPN6	36148	-5.979200375	chrX:110363997-110364459	ETS-site
CAV2	-30621	6.562976031	chr7:115895692-115896355	microsatellite
CAV2	23653	6.562976031	chr7:115950192-115950880	microsatellite
CCND1	-18841	3.751864232	chr11:69145849-69146498	microsatellite
CCND1	-41439	3.751864232	chr11:69123442-69123958	ETS-site
CD59	45687	-8.935169548	chr11:33668554-33670265	ETS-site
CD59	-31520	-8.935169548	chr11:33745985-33746596	ETS-site
CDC42BPA	-39041	-3.708542817	chr1:225611341-225611640	ETS-site
CHST12	-26573	-2.053606422	chr7:2248229-2248500	ETS-site
CIB1	46788	-2.230509354	chr15:88531177-88531755	ETS-site
CLDN1	-16834	2.008244635	chr3:191539374-191539993	ETS-site
CLDN1	-42915	2.008244635	chr3:191565652-191566024	ETS-site
CLEC11A	15473	2.362566516	chr19:55933752-55934360	microsatellite
COL3A1	27772	-3.463753053	chr2:189574960-189575328	ETS-site
COL6A2	-48347	-26.47475891	chr21:46293940-46294365	ETS-site
CPB2	-11829	2.752450593	chr13:45588860-45589672	microsatellite
CRABP2	23796	-2.976598849	chr1:154917915-154918586	ETS-site
CREB3L1	-8592	-3.158408939	chr11:46246987-46247412	ETS-site
CREB3L1	-3734	-3.158408939	chr11:46251914-46252322	ETS-site
CRIP2	22757	-2.302972313	chr14:105034865-105035103	ETS-site
DAPK1	-95333	11.03503844	chr9:89207102-89208221	microsatellite
DAPK1	-15282	11.03503844	chr9:89287171-89287875	microsatellite
DAPK1	-6101	11.03503844	chr9:89295693-89296757	ETS-site
DCLRE1A	-131528	5.680697194	chr10:115735163-115736021	microsatellite
DDAH1	26680	-3.55244668	chr1:85676615-85676881	ETS-site
DLG2	-36582	5.186261801	chr11:85052358-85052744	ETS-site
DSTN	-49167	-2.363518546	chr20:17449140-17449605	ETS-site
DUSP6	-37951	-3.252399554	chr12:88307893-88309032	ETS-site
DUSP6	5719	-3.252399554	chr12:88264253-88265041	ETS-site
ECE1	-35078	-2.121576455	chr1:21579512-21579862	ETS-site
ECE1	-45600	-2.121576455	chr1:21589992-21590703	ETS-site
ENG	-28105	-6.318341588	chr9:129684745-129685381	ETS-site
EPB41L2	-61785	3.984885798	chr6:131487531-131487965	microsatellite
EPHX1	45677	-3.042835093	chr1:224109941-224110599	ETS-site
EXOSC7	-13748	5.176423784	chr3:44978671-44979573	microsatellite
EXT2	35435	-2.828547045	chr11:44108222-44109310	ETS-site
EXT2	-29382	-2.828547045	chr11:44044044-44044562	ETS-site
F2R	-31536	-6.989317501	chr5:76015669-76016193	ETS-site
F2R	-19532	-6.989317501	chr5:76027896-76028210	ETS-site
FAM127A	-39610	-2.886942618	chrX:133954280-133954589	ETS-site
FAM127A	-18548	-2.886942618	chrX:133975286-133975611	ETS-site
FAM127A	1009	-2.886942618	chrX:133994692-133995171	ETS-site
FAM49A	-19844	19.13492211	chr2:16730106-16730706	microsatellite
FAS	-2763	3.458695416	chr10:90736012-90736776	ETS-site
FCGRT	-951	7.45909921	chr19:54706411-54707094	microsatellite

FDX1	-74833	5.572888497	chr11:109730494-109731135	microsatellite
FDX1	-31811	5.572888497	chr11:109773637-109774352	ETS-site
FNDC3B	22209	-2.9143243	chr3:173262149-173262521	ETS-site
FOLR1	-30488	-4.655283293	chr11:71547411-71547912	ETS-site
FSTL1	15215	-4.065888339	chr3:121637106-121637518	ETS-site
GALNT7	-146511	2.019025102	chr4:174179641-174180316	microsatellite
GGA2	-47148	2.052465304	chr16:23475987-23476697	ETS-site
GNAI3	-158	2.033988556	chr1:109892219-109892828	ETS-site
GNAI3	-16133	2.033988556	chr1:109875872-109876856	ETS-site
GPR56	39023	-10.65619454	chr16:56258822-56259246	ETS-site
GRP	25456	18.61408534	chr18:55063606-55064029	microsatellite
GRP	-12466	18.61408534	chr18:55025758-55026253	ETS-site
GYG2	-63356	4.444480293	chrX:2693044-2694128	microsatellite
GYG2	-11683	4.444480293	chrX:2745065-2745326	microsatellite
H2AFY2	-11338	2.817277452	chr10:71470623-71471457	microsatellite
HIPK1	-8782	2.079311617	chr1:114194207-114194666	ETS-site
HOOK1	31685	12.12388063	chr1:60084606-60085366	microsatellite
HOOK1	-45016	12.12388063	chr1:60007597-60008305	ETS-site
HRSP12	25190	2.355660494	chr8:99173244-99173554	microsatellite
IL1RAP	8154	2.927133873	chr3:191722650-191723423	microsatellite
ITGB5	35933	-3.871698191	chr3:126052621-126053111	ETS-site
ITGB5	-45427	-3.871698191	chr3:126134093-126134536	ETS-site
JAG1	29285	-4.946545938	chr20:10573065-10573716	ETS-site
JARID2	-137873	2.411763555	chr6:15216235-15217056	microsatellite
KCNH2	-1608	-2.008689386	chr7:150307832-150308093	ETS-site
KIAA0182	-41460	2.017587045	chr16:84160902-84161214	ETS-site
LAMC1	47521	-2.856894631	chr1:181306451-181306957	ETS-site
LBH	-8443	12.91513211	chr2:30299296-30299617	microsatellite
LBH	-104150	12.91513211	chr2:30203422-30204483	microsatellite
LMO2	-38279	-4.073239609	chr11:33908418-33909096	ETS-site
LMO3	-118355	5.677730537	chr12:16770371-16770901	microsatellite
LTBP3	-15356	-2.221617911	chr11:65097485-65097830	ETS-site
MAN2A1	-78711	4.992604554	chr5:108974078-108974559	microsatellite
MAN2A1	-44366	4.992604554	chr5:109008454-109009002	ETS-site
METTL3	-146084	2.613230503	chr14:21195151-21195702	microsatellite
METTL3	-23377	2.613230503	chr14:21072577-21072963	ETS-site
MFSD1	-31056	-6.294020026	chr3:159971325-159971750	ETS-site
MMP1	37667	-2.757617039	chr11:102136042-102136637	ETS-site
NBL1	9163	-10.56924495	chr1:19851019-19851948	ETS-site
NES	-4623	-9.631826206	chr1:154917915-154918586	ETS-site
NF2	38118	-3.063724135	chr22:28367481-28367879	ETS-site
NF2	-49852	-3.063724135	chr22:28279555-28279867	ETS-site
NKX2-2	-62350	15.56652058	chr20:21504724-21505183	microsatellite
NMI	-39837	4.281768691	chr2:151894363-151895010	microsatellite
NT5E	-46355	-50.52725778	chr6:86170012-86170408	ETS-site
NUDT1	-48	2.478551204	chr7:2248229-2248500	ETS-site
PAPPA	-30457	3.063276563	chr9:117925053-117925635	microsatellite
PCSK2	-57787	27.35625399	chr20:17096800-17097156	microsatellite
PGF	-25567	-6.375144543	chr14:74517277-74518050	ETS-site
PHF11	34919	-4.592547507	chr13:49002536-49003139	ETS-site
PHLDA1	10743	-22.69624627	chr12:74700845-74701676	ETS-site
PKP1	-16870	4.279050302	chr1:199501533-199502730	microsatellite
PLOD3	-271	-2.350043125	chr7:100647474-100648338	ETS-site
PLXND1	-26178	-2.378543799	chr3:130833838-130834853	ETS-site
PLXND1	-20051	-2.378543799	chr3:130828342-130828578	ETS-site
PPP1R1A	-36378	27.03427008	chr12:53304754-53305314	microsatellite
PTGER3	22599	6.73402963	chr1:71263151-71263633	microsatellite
PTPLB	-10403	2.629551704	chr3:124796859-124797279	ETS-site
RAB11FIP1	-73380	2.283005665	chr8:37949251-37949708	microsatellite

RAC1	-123	3.034488371	chr7:6380413-6380690	ETS-site
RBM28	-21710	2.725813005	chr7:127792745-127793071	microsatellite
RCC1	-7	2.377587434	chr1:28704625-28705274	ETS-site
RCN3	-16206	2.687583727	chr19:54706411-54707094	microsatellite
RDX	-58324	2.266835939	chr11:109730494-109731135	microsatellite
RFTN1	18514	-16.52705347	chr3:16511242-16511862	ETS-site
RRAGA	6	-2.205594279	chr9:19039118-19039708	ETS-site
RRBP1	16380	-3.041626373	chr20:17594354-17594730	ETS-site
RRBP1	46247	-3.041626373	chr20:17564388-17565429	ETS-site
SALL2	-120204	8.240921642	chr14:21195151-21195702	microsatellite
SF3A3	-72341	2.020511613	chr1:38299687-38301351	microsatellite
SFRS10	-77042	2.717917956	chr3:187215371-187215733	microsatellite
SHFM1	-37894	2.042817146	chr7:96214781-96215382	microsatellite
SHFM1	-44655	2.042817146	chr7:96221612-96222137	ETS-site
SIRPA	-43610	2.072521013	chr20:1778808-1779851	ETS-site
SLC16A3	-20662	-2.472404588	chr17:77757768-77759739	ETS-site
SOSTDC1	16220	-9.93689696	chr7:16520233-16520857	ETS-site
SPRY2	35039	-6.406739777	chr13:79777917-79778248	ETS-site
SPRY2	-21149	-6.406739777	chr13:79833983-79834422	ETS-site
SRPRB	-30358	-2.244250586	chr3:134976804-134977368	ETS-site
SUPT4H1	-5816	2.524538789	chr17:53790228-53790578	ETS-site
TAGLN2	35320	-2.094030377	chr1:158126345-158127041	ETS-site
TRA@	2846	-2.149042891	chr14:22088801-22089307	ETS-site
TRAM2	27456	-24.17317238	chr6:52522211-52522562	ETS-site
UGCG	29986	-3.963460009	chr9:113728527-113729003	ETS-site
UPF3B	-33985	4.606271656	chrX:118904655-118905131	ETS-site
UPP1	19397	-2.861703516	chr7:48113995-48114363	ETS-site
USP14	-47059	2.035945535	chr18:100823-101952	ETS-site
USP33	-41	4.103793574	chr1:77997490-77998316	ETS-site

Supplementary Table 7. Peaks selected for ChIP-qPCR experiments. *ETS: peaks with ETS motif, GGAA: peaks with (GGAA)n microsatelite, negative: peaks rejected by MICSA.

Type of peak	Gene	Peak height	Distance from peak to TSS
ETS*	CCND1	6.288	-18841
	CLDN1	4.561	-42915
	F2R	4	-19532
	NUDT1_1	4	-48
	NUDT1_2	4	-48
	UPF3B	5	-33985
GGAA*	AKAP7	3.87	-20352
	DAPK1	4	-15282
	FAM49A	7.769	-19844
	GYG2	4	-11683
	NMI	5	-39837
	PAPPA	5	-30457
	PKP1	4.768	-16870
	PPP1R1A	7.787	-36378
	RCN3	6	-16206
	SLCO5A1	7.909	-36342
	TNFAIP6	5	-27850
negative*	ACADVL	5	3226
	ATF6B	4	6962
	CAD	5	18610
	CANT1	6	7382
	GNB1	5	98762
	PRDM2	4	106009
	TCF7L1	8	82772
	USP36	4	42332

Supplementary Methods

Motifs. A set of overrepresented motifs is de novo identified from the set of *N* “*confident peaks*”. The default value of *N* is three hundred. However, this parameter is optional so that the user can experiment with it in case he or she is unsatisfied with identified motifs.

The “confident peaks” are selected in the following way:

From the whole set of candidate peaks in ChIP data we select peaks of heights greater than heights in the control set. This subset usually contains much more than *N*=300 peaks. If this is the case, we select 300 peaks randomly out of this subset. Otherwise, we use the whole subset for motif selection although it contains fewer peaks.

The idea that is behind such a selection procedure is the following: in the case where the transcription factor has two motifs, the first one may correspond to strong binding and the second one to weak binding. Thus, we should not restrict ourselves to only the top peaks since by doing so we can fail to identify the second motif.

We decided to take 300 peaks for motif detection for a particular reason. Much greater values would make motif identification by MEME [1] impossible in a reasonable time. On the other hand, taking a much smaller number of peaks may not allow identification of the position weight matrix with high precision, with correct flanking regions and length.

As we believe that real binding motifs occur in the central area of peaks, we select only central areas of *N* peaks to identify motifs.

We run MEME on this set with the following parameters:

-revcomp (search on both strands)

-dna (use DNA 4 letters alphabet)

-mod zoops (expect to find motif zero or one times in a sequence)

-evt 0.5 (maximal E-value for a motif is 0.5)

-minw 7 (minimal motif length is 7)

-maxw 24 (maximal motif length is 24)

We run MEME three times to identify up to three motifs for dataset *D*, each time we keep the best identified motif.

The first run of MEME results in set *{H$_i$}* of sequences that constitute the most overrepresented motif *M$_1$* in *D*. Using the set *{H$_i$}* we construct a PSSM: $PSSM[i][\alpha] = ln\left(\frac{counts[i][\alpha] + pseudoScore * p_\alpha}{N + pseudoScore}/p_\alpha\right)$, where $counts[i][\alpha]$ is the number of sites where the nucleotide α is observed in position *i*, p_α is a background probability of nucleotide α, *N* is a number of sites, $pseudoScore = \ln(N)$ is a pseudo-score.

The set *D* of *S significant peaks* is then searched again for occurrences of *M$_1$* with its PSSM and minimal score threshold *minT.* The minimal threshold *minT* for the PSSM score is selected as $minT = minThr - \frac{4 \cdot minThr}{maxThr - minThr}, minThr = \min_i PSSM(H_i),\ maxThr = \max_i PSSM(H_i)$. Here $PSSM(H_i)$ is a score of word H_i calculated with the PSSM of motif *M$_1$*. The correction $\frac{4 \cdot minThr}{maxThr - minThr}$ is introduced to weaken strong motifs.

Sequences from *D* which do not contain such occurrences form the second set *D$_2$*. They are subject to the second run MEME. If MEME finds the second overrepresented motif *M$_2$*, the set *D$_3$* of sequences which do not contain motif *M$_2$* will be constructed. Then, MEME is run for the third time to identify the third motif if this motif exists.

Classes. Below, we consider separately each chromosome. The whole set of peaks for each chromosome is divided into classes so that each class *C$_i$* contains peaks with the same number of overlapping DNA fragments *i*. For example (see Supplementary Fig. 10), nine DNA fragments overlap, forming a peak which belongs to class *C$_9$*. We believe that all peaks belonging to the same class have the same properties; below, we use the same statistical parameters for all peaks from the same class.

Initial FDR. Since we have two datasets: one from ChIP and the other from the control experiment, we can calculate the *initial false discovery rate* (FDR) for each class *C$_i$*, which is a ratio:

$inFDR_i = \frac{K_i}{M_i}$, where *K$_i$* is the number of peaks with depth of DNA fragment coverage equal to or greater than *i* in the control dataset, and *M$_i$* is the number of peaks with the depth of DNA fragment coverage equal to or greater than *i* in the ChIP dataset.

The FDR value represents the probability that a peak in $\{C_j\}_{j\geq i}$ is a result of random events and is not a real binding site.

Optimization. The last step of the MICSA pipeline is the optimization procedure. It aims to maximize the total number of reported peaks for each class so that the total number of expected false positives does not exceed the user-defined threshold (*F*). Below, we focus the explanation on the case of one motif *M* and one chromosome.

For each class of peaks C_i, we optimize the following parameters: $T_{M,i}$ which is a PSSM score threshold for a given motif *M*, and $\delta_{M,i}$ which is a parameter specifying the length of the central area of the peak (Supplementary Fig. 10).

For each class C_i and each pair ($T_{M,i}$, $\delta_{M,i}$) we can calculate $S_i(T_{M,i}, \delta_{M,i})$ which is the number of peaks that we select when we only keep peaks with at least one occurrence of motif *M* (PSSM score ≥ $T_{M,i}$) in the central area specified by $\delta_{M,i}$. Independently, we can estimate the number of false peaks selected by chance using the procedure:

$$F_i\left(T_{M,i}\,, \delta_{M,i}\right) = \sum_{peak\ j\ in\ C_i} (inFDR_i)\left(mp_j\right),$$

where $inFDR_i$ is the *initial FDR* of class C_i defined above and mp_j is a motif p-value of peak *j* described in the following.

Each term in the sum is the probability of two events considered to be independent: that a peak has DNA fragment coverage equal to or greater than *i* just by chance and that by chance this peak contains a motif occurrence.

The *motif p-value* for peak *j* (mp_j) is a probability to observe a motif by chance in a sequence of given length. In our case the motif is represented by its PSSM with the threshold $T_{M,i}$, and the sequence length $L(\delta_{M,i}, j)$ is equal to the length of central area of peak *j* provided by $\delta_{M,i}$. We use the Poisson approximation to calculate the motif p-value: $motif\ p\text{-}value \approx 1 - (1 - P(M))^{L(\delta_{M,i},j) - MotifLength + 1}$. Here, the *motif probability* *P*(*M*) is the probability of observing a motif occurrence with a PSSM score above a given threshold $T_{M,i}$ at a given position. We consider it equal to the genomic frequency of the motif with threshold $T_{M,i}$. Since it is very time-consuming to evaluate motif frequencies in a whole genome, in our approach we use only chromosome 1.

The aim of the optimization is to find such values of ($T_{M,i}$, $\delta_{M,i}$) that maximize $\Sigma S_i(T_{M,i}, \delta_{M,i})$ so that $\Sigma F_i(T_{M,i}, \delta_{M,i})$ stays below the user-defined threshold *F*. Since it is very time consuming to do it exactly, we discretize all parameters.

We divide segment]0, *F*] into very small sub-segments:]0, *ε*],]*ε*, *2ε*],… ,]F-*2ε*, *F*- *ε*],]F- *ε*, *F*]. We set $\varepsilon = F/1000$. This gives 1000 sub-fragments. For the first class C_1 and for each segment]*kε*, (*k+1*)*ε*] we find such ($T_{M,1}$, $\delta_{M,1}$) so that $k\varepsilon < F_1(T_{M,i}, \delta_{M,i}) \leq (k+1)\varepsilon$ and $S_1(T_{M,1}, \delta_{M,1})$ are maximal. We fill array B_1 with corresponding maximized values of $S_1(T_{M,1}, \delta_{M,1})$. We remember the choice of ($T_{M,1}$, δ_{M1}) for each sub-fragment. We create an array *A* with 1000 elements that should contain corresponding values of $S_i(T_{M,i}, \delta_{M,i})$ at each following step (this array is called A_i in step *i*). In the first step we fill array $A_1$with values from B_1: $A_1[j] = B_1[j]$. Then, for each following class C_i we repeat the same procedure, i.e., for each segment]*kε*, (*k+1*)*ε*] we find such ($T_{M,i}$, δ_{Mi}) so that $k\varepsilon < F_i(T_{M,i}, \delta_{M,i}) \leq (k+1)\varepsilon$ and $S_i(T_{M,i}, \delta_{M,i})$ would be maximal. Corresponding maximized values of $S_i(T_{M,i}, \delta_{M,i})$ are held in array B_i. Then, we fill A_i as follows: $A_i[j] = \max_{k=\overline{1,j}}(A_{i-1}(k) + B_i(j-k))$. After the last step i_{last}, the value $A_{i_{last}}$,$[1000]$ will contain a value close to $\max\left(\Sigma S_i(T_{M,i}, \delta_{M,i})\middle|\, F_i\left(T_{M,i}, \delta_{M,i}\right) \leq F\right)$.

This procedure guarantees that at the end we will find such ($T_{M,i}$, $\delta_{M,i}$), so that

$$F - \varepsilon \cdot NumberOfClasses \leq \Sigma F_i\left(T_{M,i}, \delta_{M,i}\right) \leq \mathrm{F}, \text{ and}$$

$$\Sigma S_i\left(T_{M,i}, \delta_{M,i}\right) \geq (\text{real}\max \Sigma S_i \mid \Sigma F_i \leq \mathrm{F} - \varepsilon \cdot NumberOfClasses).$$

Different values of *F* result in different numbers of reported peaks for each class C_i (Supplementary Fig. 2).

The user can choose between setting the total maximum value of false positive peaks (*F*) or setting the maximum ratio between expected number of false positives and the number of called peaks. This ratio is called false discovery rate: FDR = $\mathrm{F}/\Sigma S_i$.

Program execution. The tutorial on how to use the graphical interface of MICSA (Supplementary Fig. 11) can be found on the MICSA website http://bioinfo-out.curie.fr/projects/micsa/tutorial.html.

Below are instructions on how to run the MICSA pipeline from the command line.

1. Run FindPeaks for ChIP and control data:

```
java -Xmx2G -jar FindPeaks.jar -aligner <aligner> -eff_frac <eff_frac> -duplicatefilter -input <ChIP input
files> -name chip -output <existing output directory> -dist_type <dist_type> -minimum 3
```

```
java -Xmx2G -jar FindPeaks.jar -aligner <aligner> -eff_frac <eff_frac> -duplicatefilter -input <control
input files> -name control -output <existing output directory> -dist_type <dist_type> -minimum 1
```

More details on FindPeaks parameters can be found at http://vancouvershortr.wiki.sourceforge.net/FindPeaks4

2. Filter out peaks in repetitive or ambiguous regions of genome:

```
java DeleteRegions -f chip_triangle_standard.peaks -r <file with positions to mask>
```

```
java deleteRegions -f control_triangle_standard.peaks -r <file with positions to mask>
```

3. Create summary about peak distribution in ChIP and control data:

```
java Summary -f chip_triangle_standard.peaks -c control_triangle_standard.peaks -r <ratio>
```

4. Filter out peaks occurring both in ChIP and Control data:

```
java FilterPeaks -f chip_triangle_standard.peaks -c control_triangle_standard.peaks -t
<coverage_threshold>
```

Supplementary Figure 12 describes the background filtering procedure.

5. Run MICSA.jar :

```
java -jar -Xmx2G micsa.jar -name <name> -f chip_triangle_standard.peaks [-n <max_false_positive> -
fdr <FDR value>] -o <output_dir> -l <file with summary> -g <genome_dir> -w <wig_file>
```

See the MICSA tutorial http://bioinfo-out.curie.fr/projects/micsa/tutorial.html for more details on parameters and an example of MICSA run on the NRSF dataset [2].

References.

1. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.
2. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.