Supplementary notes

Influence of the background model on the motif analysis

Summary:

We show here that the under-representation of the AT-rich motifs in the peak centers is not due to an inappropriate background model. We base this conclusion on several additional experiments we performed. First, using the same motif-scanning approach as in the main study, we tested several background models of various Markov orders, and local background models in sliding windows, to avoid over-fitting conditions. Furthermore, we show that the background model used recapitulates the specific composition of the peak dataset, and that the depletion is still detectable with these various background models.

As a complement, we used motif-discovery approaches: one that specifically looks for positional biases without explicit background model, and another approach to identify globally under-represented motifs. Here again, we still detect the local under-representation of AT-rich sequences.

We also performed an analyses on DNaseI hypersensitive sites derived from ENCODE data, which show the same local depletion of AT-rich sequences arguing that AT-depletion is a general feature of genomic regions involved in gene regulation.

A – Datasets

The dataset used for this test is the ChIP-seq peaks in U2OS cells (ArrayExpress accession:E-MTAB-2731) (Fig. A1 & A2).

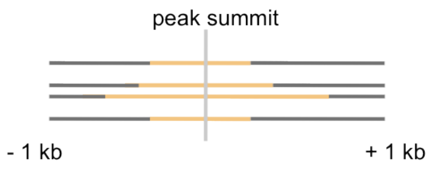
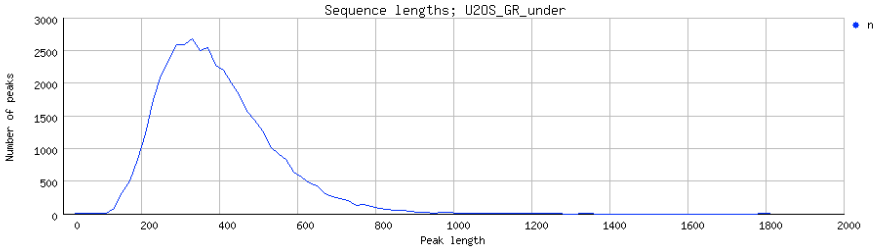


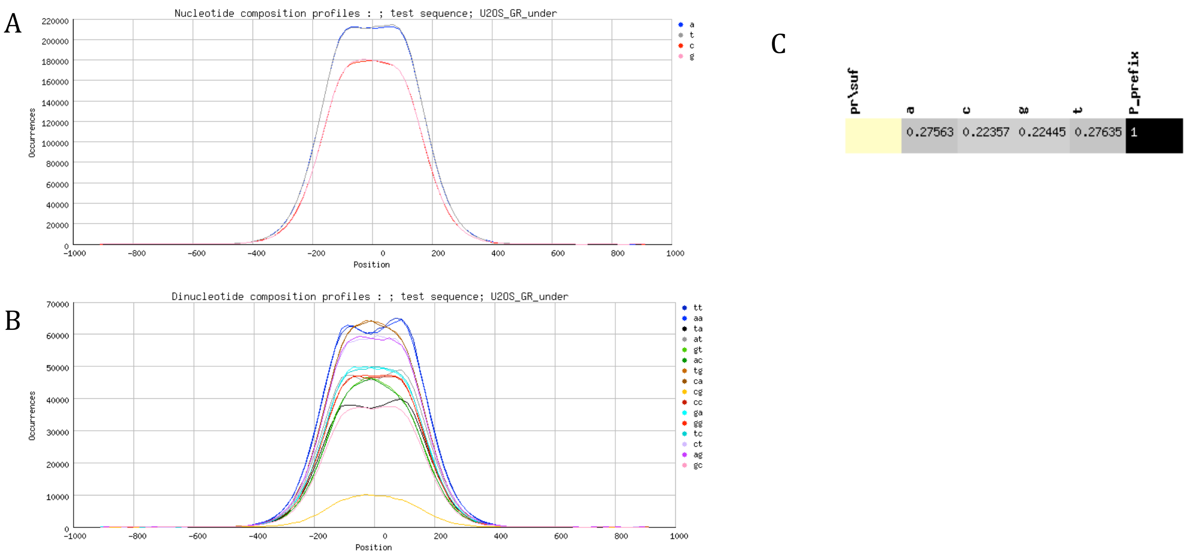
Figure A1: the “peak” dataset corresponds to the yellow fragments. The “2kb” dataset corresponds to the regions +/-1kb centered on the peak summit.



Nb of peaks: 41402
Total seq. size: 16224 kb
Min length: 20 bp
Mean length: 391.864 bp
Max length: 1810 bp

Figure A2: Lengths of sequences in the “peak” dataset, and dataset statistics.

The composition profile of the 41402 peaks reveals a higher number of A and T nucleotides (Fig. A3). The dinucleotide profile highlights very low CpG occurences, typical of mammalian sequences [1]. For AA and TT (blue), which show the highest occurrences, we observe a local decrease at the center of the peaks. This local decrease is also detectable for TA (black) and AT (grey) even if these dinucleotides have lower occurrences overall. Note that the global nucleotide composition is roughly the same for each nucleotide (Fig. A3c).

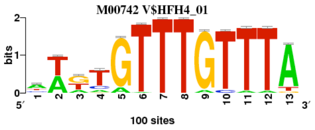


pr\suf	a	c	g	t	P_prefix
	0.27563	0.22357	0.22445	0.27635	1

Figure A3: A: nucleotide composition profile. **B**: dinucleotide composition profile. **C**: Global composition of the peaks

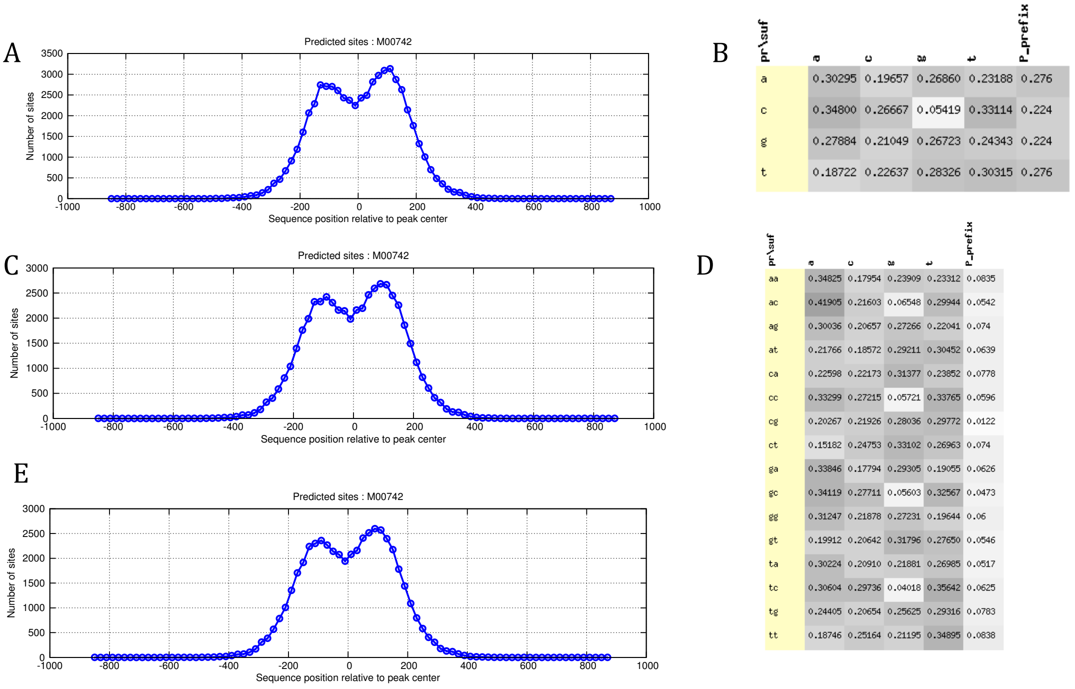
B – Motif scanning

The analyses of this section focus on motif scanning, with a PSSM (here Transfac M00742 – HFH4) that is under-represented in our study (Figure 1 of the main text) and is AT-rich (%GC: 0.22). We test the effect of the background model used for scanning, which most generally is a markov 0 (Bernoulli, mononucleotide) as most motif-scanning programs only supports Bernoulli models (e.g. FIMO from MEME suite). We used the RSAT *matrix-scan* program that allows higher-order markov models, and tested various order and various background datasets. Note that it is not possible to train a high-order background on a small dataset, to prevent over-fitting [2]. Increasing the order drastically increases the computing time.



1) Peak dataset (not repeat-masked)

Focusing on the ChIP-seq peaks with a background model calculated on these same peaks, we observe a local depletion of the HFH4 motif in the center of the peaks, even with a high-order background model, taking into account up to 4-mers frequencies (Fig. B1A,C,E). The background model correctly captures the particularities of the dataset (low CpG, high frequency of AA and TT (Fig B1B,D).



pr\suf	a	c	g	t	P_prefix
a	0.30295	0.19657	0.26860	0.23188	0.276
c	0.34800	0.26667	0.05419	0.33114	0.224
g	0.27884	0.21049	0.26723	0.24343	0.224
t	0.18722	0.22637	0.28326	0.30315	0.276

pr\suf	a	c	g	t	P_prefix
aa	0.34825	0.17954	0.23909	0.23312	0.0835
ac	0.41905	0.21603	0.06548	0.29944	0.0542
ag	0.30036	0.20657	0.27266	0.22041	0.074
at	0.21766	0.18572	0.29211	0.30452	0.0639
ca	0.22598	0.22173	0.31377	0.23852	0.0778
cc	0.33299	0.27215	0.05721	0.33765	0.0596
cg	0.20267	0.21926	0.28036	0.29772	0.0122
ct	0.15182	0.24753	0.33102	0.26963	0.074
ga	0.33846	0.17794	0.29305	0.19055	0.0626
gc	0.34119	0.27711	0.05603	0.32567	0.0473
gg	0.31247	0.21878	0.27231	0.19644	0.06
gt	0.19912	0.20642	0.31796	0.27650	0.0546
ta	0.30224	0.20910	0.21881	0.26985	0.0517
tc	0.30604	0.29736	0.04018	0.35642	0.0625
tg	0.24405	0.20654	0.25625	0.29316	0.0783
tt	0.18746	0.25164	0.21195	0.34895	0.0838

Figure B1: Motif profile on the peaks for matrix M00742 with background models calculated on the peaks. **A**: markov1 (dinucleotides). **B**: background model markov 1. **C**: markov2 (trinucleotides). **D**: background model markov 2. **E.** markov3 (4-mers)

2) 2kb dataset

a. peak background model

This is the model used in the main study, which shows an under-representation of several AT-rich (and few non-AT rich) motifs (Figure 1).

b. local window background model

To account for the local variations in nucleotide composition, we tested local background models with RSAT *matrix-scan*. A window is defined around the PSSM, in which the background model is calculated (excluding the sequence at the center, aligned with the PSSM). The window slides by 1bp, over each sequence. A very local background model can only be calculated for low order Markov models (to prevent over-fitting). We tested Markov 0 over in 75bp window (Fig.B2A) and Markov 1 in 200bp window (Fig.B2B). As these calculations are time-consuming (the background model is recalculated at each position of each sequences of the dataset, and the theoretical distribution of scores must be recalculated as well to compute the p-values), using a cluster was necessary even for just one matrix. Even with a local background, which takes the dinucleotide composition into account, we still observe the local depletion at the center. The number of hits is nevertheless lower (compare the Y axis between panels A an B with figure 1).

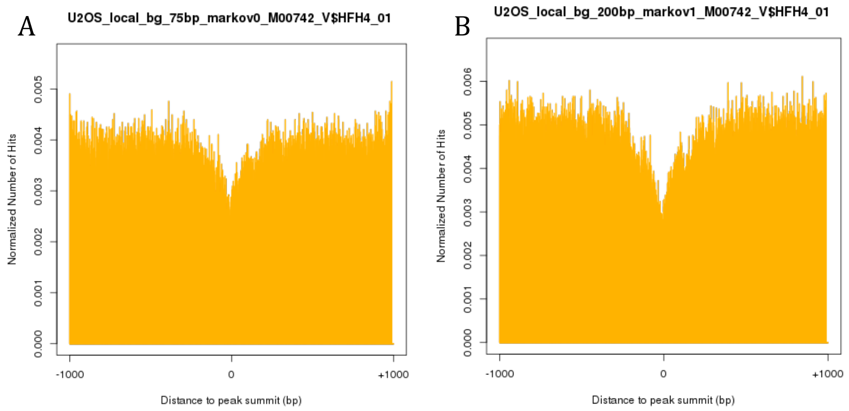


Figure B2: Motif profile on the 2kb dataset for matrix M00742 with local background models in sliding windows. **A**: 75bp window, Markov 0 (mononucleotides) **B**: 200bp window, Markov 1

C – Motif discovery

The analyses of this section focus on motif discovery, without *a priori* knowledge of the searched motifs (contrary to motif scanning, there are no PSSMs).

1) Positional bias (no explicit background model)

The program RSAT *position-analysis* does not use an explicit background model, but finds exceptional k-mers with a bias in their position [1]. Applying this approach on the 2kb dataset reveals a local depletion of k-mers at the center of the dataset (within the peaks) (Fig.C1a). As repeats are usually excluded from the peaks due to the ChIP-seq processing (the read mapping step is usually done excluding the multi-mapping reads, hence excluding repeat regions), we also performed the same analysis on repeat-masked sequences (Fig.C1b). The results are similar, with a slight decrease in significance values. Although the most significantly under-represented 6-mers are AT-rich sequences without obvious specificity, some under-represented 6-mers are not AT-rich (Fig. C2).

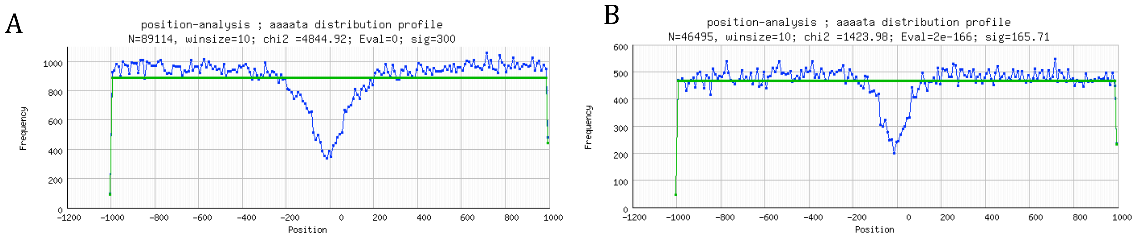


Figure C1: **A**: Local under-representation of the motif AAAATA in the 2kb dataset. **B**: same with repeat-masked sequences.

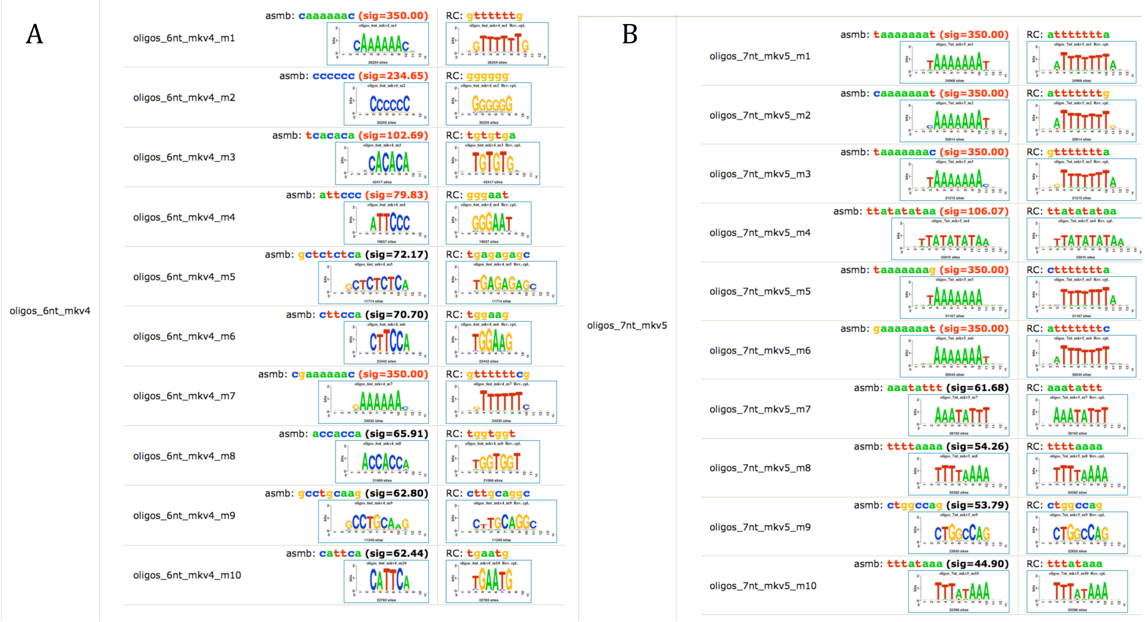
<table>
<tr><th>Sequence</th><th>chi_value</th><th></th></tr>
<tr><td>AAAAAT</td><td>6546.3</td><td>AAAAAT|ATTTTT</td></tr>
<tr><td>AAAAAA</td><td>5296.7</td><td>AAAAAA|TTTTTT</td></tr>
<tr><td>AAAATA</td><td>4844.9</td><td>AAAATA|TATTTT</td></tr>
<tr><td>AAAATT</td><td>4504.9</td><td>AAAATT|AATTTT</td></tr>
<tr><td>TAAAAA</td><td>3880.9</td><td>TAAAAA|TTTTTA</td></tr>
<tr><td>AAATTA</td><td>3831</td><td>AAATTA|TAATTT</td></tr>
<tr><td>CAAAAA</td><td>3195.1</td><td>CAAAAA|TTTTTG</td></tr>
<tr><td>TTAAAA</td><td>3117.6</td><td>TTAAAA|TTTTAA</td></tr>
<tr><td>AATATA</td><td>2784</td><td>AATATA|TATATT</td></tr>
<tr><td>**GCCTCC**</td><td>**2741.4**</td><td>**GCCTCC|GGAGGC**</td></tr>
<tr><td>GGATTA</td><td>2671.8</td><td>GGATTA|TAATCC</td></tr>
<tr><td>ATATAT</td><td>2611.5</td><td>ATATAT|ATATAT</td></tr>
<tr><td>**AGGCTG**</td><td>**2564.8**</td><td>**AGGCTG|CAGCCT**</td></tr>
<tr><td>AAATAA</td><td>2532.9</td><td>AAATAA|TTATTT</td></tr>
<tr><td>ATTTTA</td><td>2496.8</td><td>ATTTTA|TAAAAT</td></tr>
<tr><td>AAATAT</td><td>2440</td><td>AAATAT|ATATTT</td></tr>
<tr><td>GATTAC</td><td>2398.9</td><td>GATTAC|GTAATC</td></tr>
<tr><td>TATATA</td><td>2280.1</td><td>TATATA|TATATA</td></tr>
</table>

Figure C2: k-mers with the highest significance value (300). The non AT-rich k-mers are indicated in bold.

2) Global under-represented motifs

a. Peak-motifs on peak dataset

We added the possibility in the RSAT program *peak-motifs* to search for overall under-represented k-mers (only with the algorithm *oligo-analysis*). For 6-mers, we find various under represented motifs (Fig.C3A), some of which showing low complexity. For 7-mers, we find almost exclusively AT-rich motifs (Fig. C3B).



oligos_6nt_mkv4_m1	asmb: caaaaaac (sig=350.00)	RC: gtttttttg
oligos_6nt_mkv4_m2	asmb: cccccc (sig=234.65)	RC: gggggg
oligos_6nt_mkv4_m3	asmb: tcacaca (sig=102.69)	RC: tgtgtga
oligos_6nt_mkv4_m4	asmb: attccc (sig=79.83)	RC: gggaat
oligos_6nt_mkv4_m5	asmb: gctctctca (sig=72.17)	RC: tgagagagc
oligos_6nt_mkv4_m6	asmb: cttcca (sig=70.70)	RC: tggaag
oligos_6nt_mkv4_m7	asmb: cgaaaaaac (sig=350.00)	RC: gtttttttcg
oligos_6nt_mkv4_m8	asmb: accacca (sig=65.91)	RC: tggtggt
oligos_6nt_mkv4_m9	asmb: gcctgcaag (sig=62.80)	RC: cttgcaggc
oligos_6nt_mkv4_m10	asmb: cattca (sig=62.44)	RC: tgaatg

oligos_7nt_mkv5_m1	asmb: taaaaaaat (sig=350.00)	RC: atttttttta
oligos_7nt_mkv5_m2	asmb: caaaaaaat (sig=350.00)	RC: attttttttg
oligos_7nt_mkv5_m3	asmb: taaaaaaac (sig=350.00)	RC: gttttttta
oligos_7nt_mkv5_m4	asmb: ttatatataa (sig=106.07)	RC: ttatatataa
oligos_7nt_mkv5_m5	asmb: taaaaaaag (sig=350.00)	RC: cttttttta
oligos_7nt_mkv5_m6	asmb: gaaaaaaat (sig=350.00)	RC: atttttttc
oligos_7nt_mkv5_m7	asmb: aaatattt (sig=61.68)	RC: aaatattt
oligos_7nt_mkv5_m8	asmb: ttttaaaa (sig=54.26)	RC: ttttaaaa
oligos_7nt_mkv5_m9	asmb: ctggccag (sig=53.79)	RC: ctggccag
oligos_7nt_mkv5_m10	asmb: tttataaa (sig=44.90)	RC: tttataaa

Figure C3: peak-motifs analysis for globally under-represented motifs **A**: 6-mers (automatic background model markov4). **B**: 7-mers (automatic background model markov5). The background models are theoretical markov models calculated on the peak dataset.

b. Peak-motifs with two datasets: peaks and not-peaks

As control, we prepared a dataset with the same length/number of sequences as the peak dataset, but taken from genomic regions that are not located under a GR ChIP-seq peak in any of the cell lines examined (U2OS, A549, IMR90, K562, Nalm6), nor in the ENCODE blacklist. The dinucleotide profile does not feature the local decrease of AT-rich sequences (Fig.C4A). Consequently, the main difference found are these AT-rich regions in the center of the peaks (Fig.C4B).

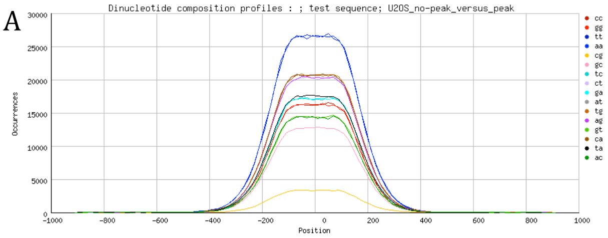


Figure C4: peak-motifs analysis for under-represented motifs between the peak dataset and the control dataset. The background model is an empirical background model calculated on the peak dataset, to search for under-represented motifs. **A**: dinucleotide composition profile of the control dataset.

c. Peak-motifs on DNaseI dataset

We randomly extracted the same number of sequences from the DNaseI ENCODE dataset (wgEncodeRegDnaseClusteredV3) and performed a *peak-motifs* analysis searching for under-representation. We first observe that the dinucleotide profile showed a similar localized depletion of AA and TA as the one we saw for the GR-bound regions. The most significant under-represented motifs are AT-rich sequences, again resembling what we found for GR-bound peaks.

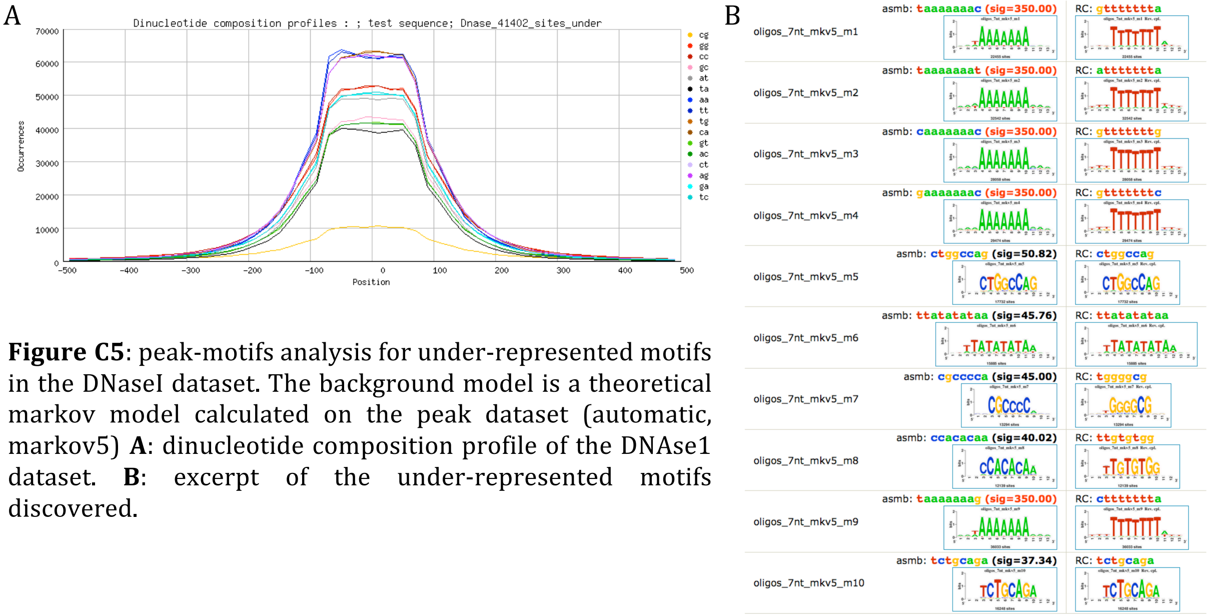


Figure C5: peak-motifs analysis for under-represented motifs in the DNaseI dataset. The background model is a theoretical markov model calculated on the peak dataset (automatic, markov5) **A**: dinucleotide composition profile of the DNAse1 dataset. **B**: excerpt of the under-represented motifs discovered.

References

[1] Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols* **7**, 1551–1568.

[2] Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols* **3**, 1578–1588.