Supplementary Data for:

Bind-n-Seq: high-throughput analysis of *in vitro* protein-DNA interactions using massively parallel sequencing

Artem Zykovich, Ian Korf*, and David J Segal*

Genome Center, University of California, Davis, CA, 95616.

***Correspondence should be addressed to:**
D.J.S. (djsegal@ucdavis.edu) and I.K. (ifkorf@ucdavis.edu)

Supplementary Table S1A. Bar codes for Bind-n-Seq run 1.

Bar Code	Protein	Protein (nM)	Salt (mM)	Extra
AAA	No-protein	0	100	
AAC	Aart	50	1	
AAG	Aart	50	10	
AAT	Aart	50	50	
ACA	Aart	50	100	
ACC	Aart	50	500	
ACG	Aart	5	100	
ACT	Aart	0.5	100	
AGA	Aart	0.05	100	
AGC	Aart	50	100	Gel shift
AGG	No-protein	0	100	
AGT	Zif268	50	1	
ATA	Zif268	50	10	
ATC	Zif268	50	50	
ATG	Zif268	50	100	
ATT	Zif268	50	500	
CAA	Zif268	5	100	
CAC	Zif268	0.5	100	
CAG	Zif268	0.05	100	
CAT	Zif268	50	100	Gel shift
CCA	Oligo only			Sigma
CCC	Oligo only			Sigma

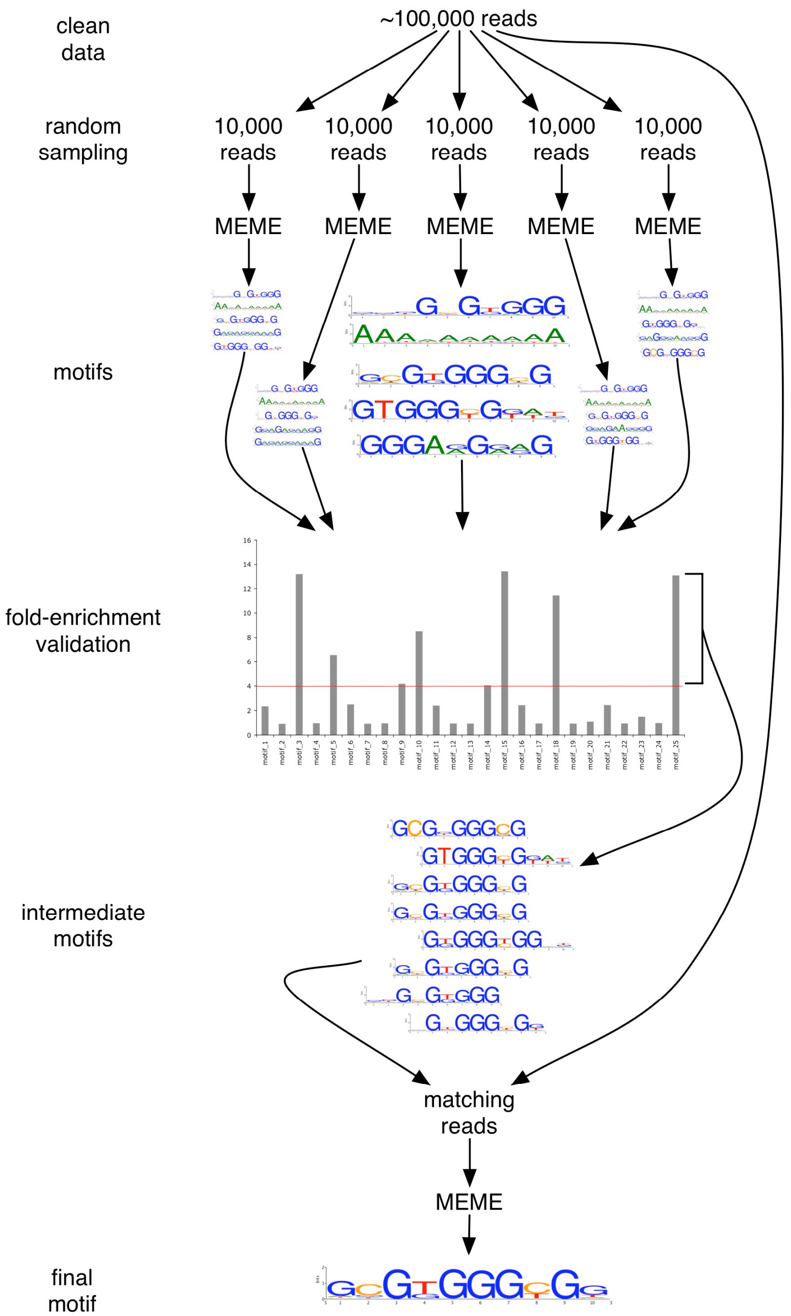
Supplementary Table S1B. Bar codes for Bind-n-Seq run 2.

Bar Code	**Protein**	**Protein (nM)**	**Salt (mM)**	**Extra**
AAA	No-protein	0	100	
AAC	Zif268	50	1	
AAG	Zif268	50	10	
AAT	Zif268	50	50	
ACA	Zif268	50	100	
ACC	Zif268	50	500	
ACG	Zif268	5	100	
ACT	Zif268	0.5	100	
AGA	Zif268	0.05	100	
AGC	Zif268	50	100	Gel shift
AGG	No-protein	0	100	
AGT	Aart	50	1	
ATA	Aart	50	10	
ATC	Aart	50	50	
ATG	Aart	50	100	
ATT	Aart	50	500	
CAA	Aart	5	100	
CAC	Aart	0.5	100	
CAG	Aart	0.05	100	
CAT	Aart	50	100	Gel shift
CCA	Oligo only			Sigma
CCC	Oligo only			Sigma
CCG	Zif268	5	100	Long wash
CCT	Zif268	5	100	+ round
CGA	Zif268	5	200	
CGC	Zif268	5	100	Ficoll
CGT	Aart	5	100	Long wash
CTA	Aart	5	100	+ round
CTC	Aart	5	200	
CTG	Aart	5	100	Ficoll
TGG	Oligo only			Sigma hand mix
TGT	Oligo only			Sigma hand mix
TTA	Oligo only			IDT
TTC	Oligo only			IDT
TTG	Oligo only			Bioneer
TTT	Oligo only			Bioneer

Supplementary Table S2. Relative affinity by QuMFRA from Liu et al, 2005

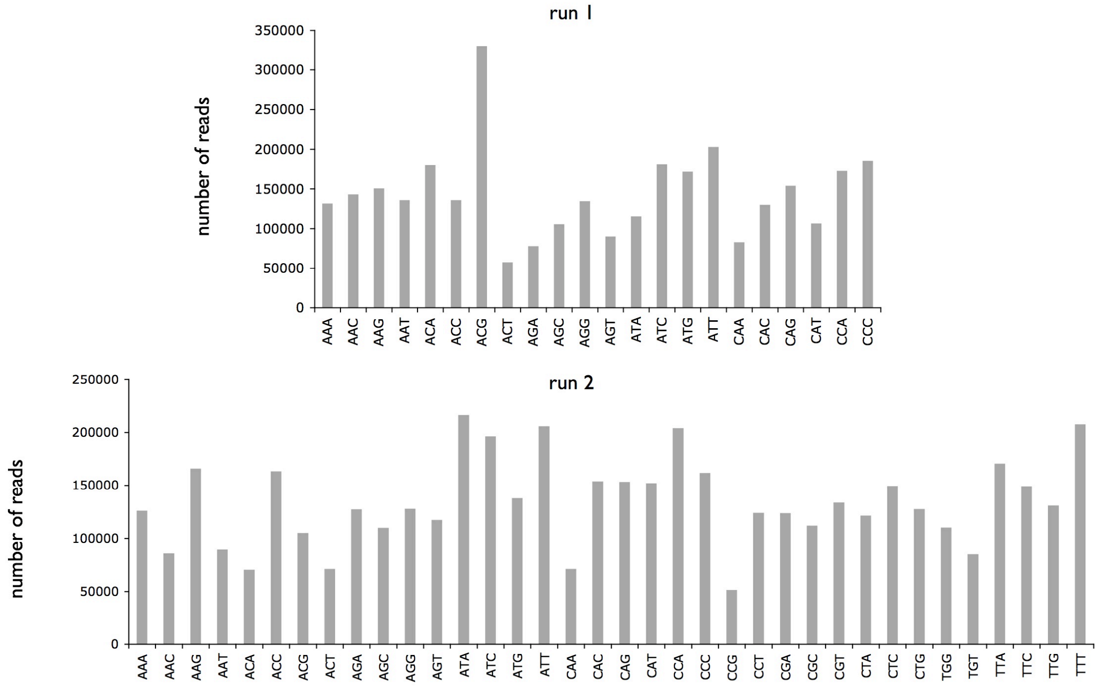
Sequence	Relative Ka
GCGTGGGCGT	39.93
GCGTGGGCGG	23.32
GCGTGGGAGG	10.36
GCGTGGGCAT	7.22
GCGTGGGTTT	5.68
GCGTGGGGTG	4.58
GCGTGGTGTG	4.01
GCGTGGGGGA	2.14
GCGTGGTGCG	1.78
GCGTGGCCGT	1.19
GCGTGGGGTA	1.17
GCGTGGGTGC	1.15
GCGTGGGGGT	1.00
GCGTGGTGAG	0.39
GCGTGGGATC	0.04

Supplementary Figure S1: De novo motif finding**.** An overview of the bioinformatics processing showing reads that have been sorted, filtered (clean data), and split into 5 random, non-overlapping subsets of 10,000 reads for motif finding with MEME. The best 5 motifs from each run of MEME are scored against protein-containing and no-protein reads to find motifs with the greatest fold-enrichment. Motifs with 4-fold or greater enrichment are kept as the intermediate motifs. Reads matching the intermediate motifs are run through MEME again to produce the final motif(s).

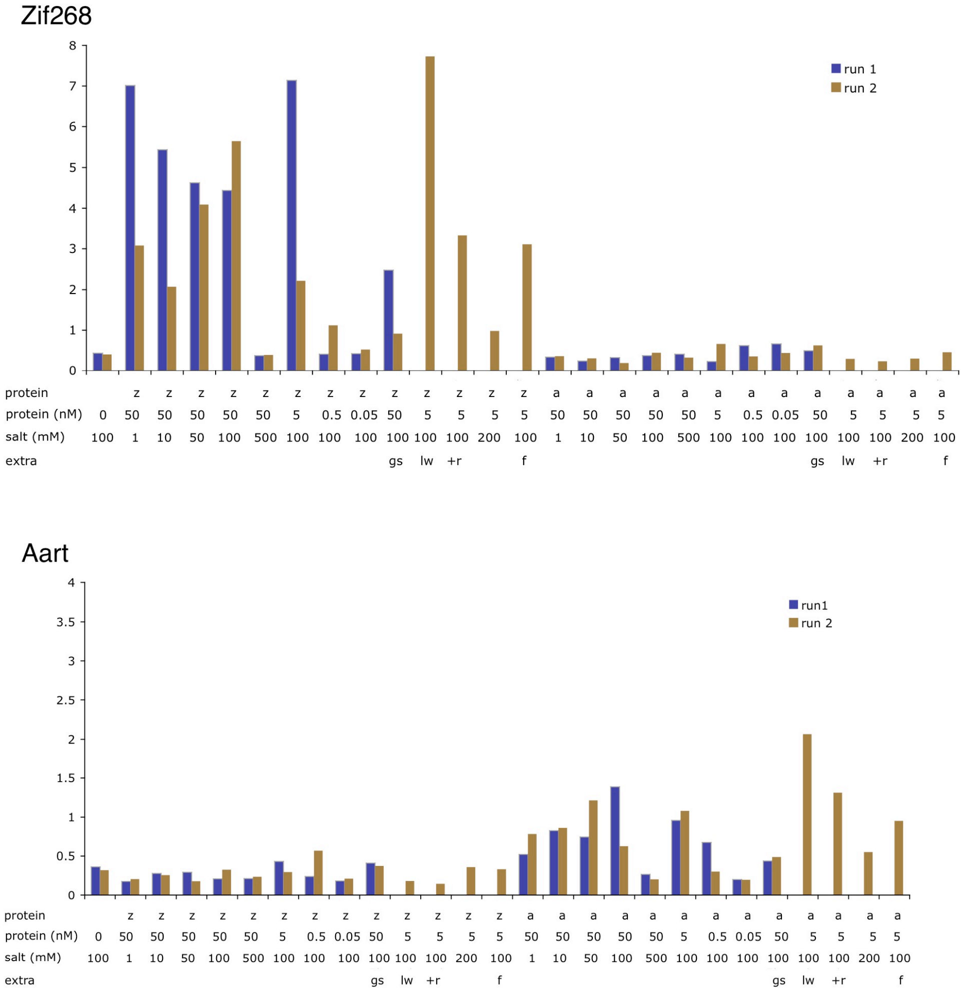


Supplementary Figure S2: Sorting sequences by bar codes.

	run 1	run 2
Number of reads (raw data)	7275683	9390930
Number of reads (clean data)	3327661	5086559

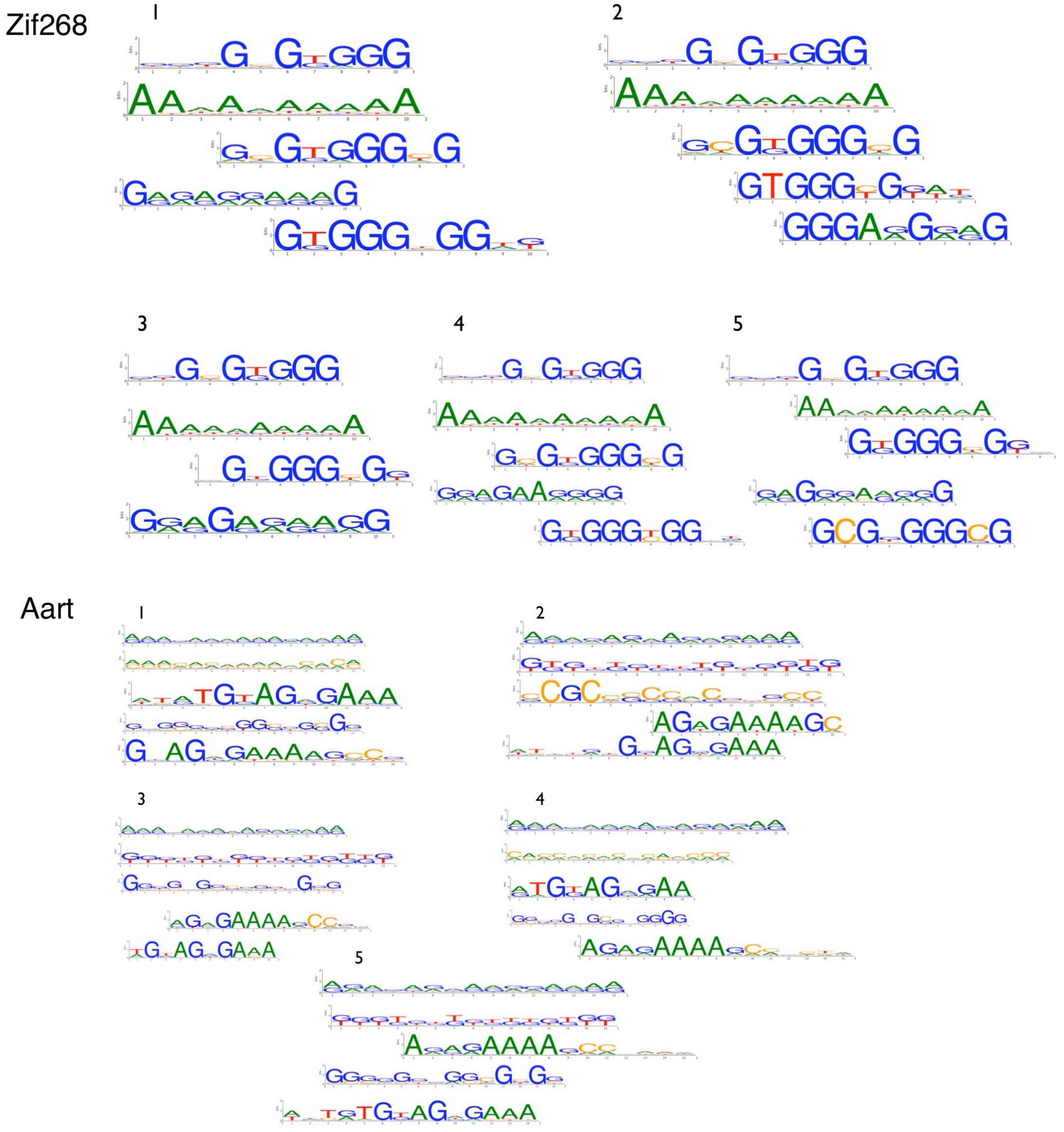


Supplementary Figure S3: Percentage of the expected motif found in "clean" data.

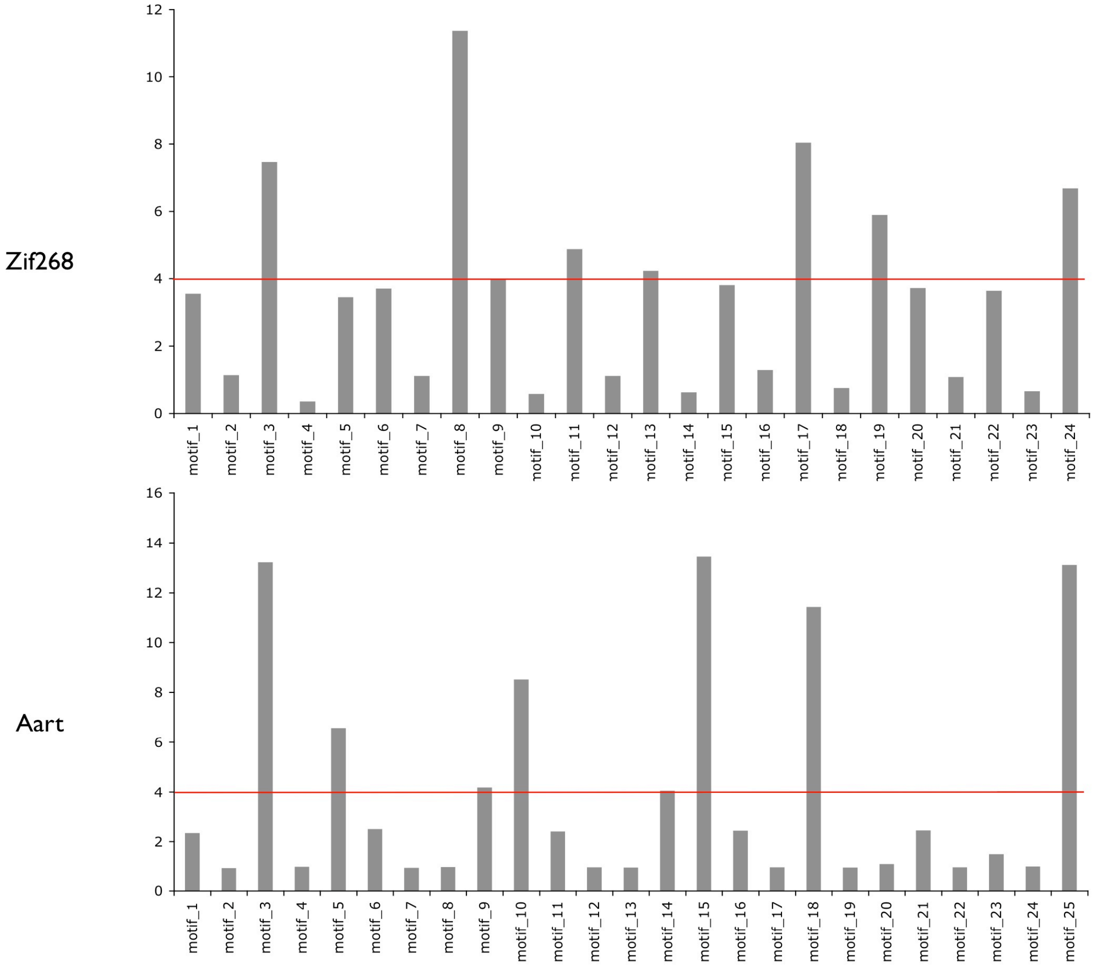

Zif268

Aart

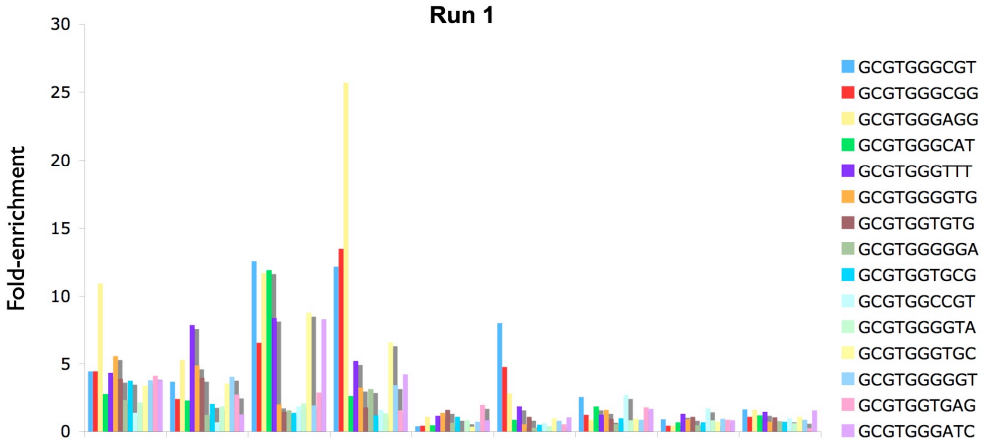
Supplementary Figure S4: The five best motifs. For Zif268 and Aart, each group of motifs was found from 10,000 reads.



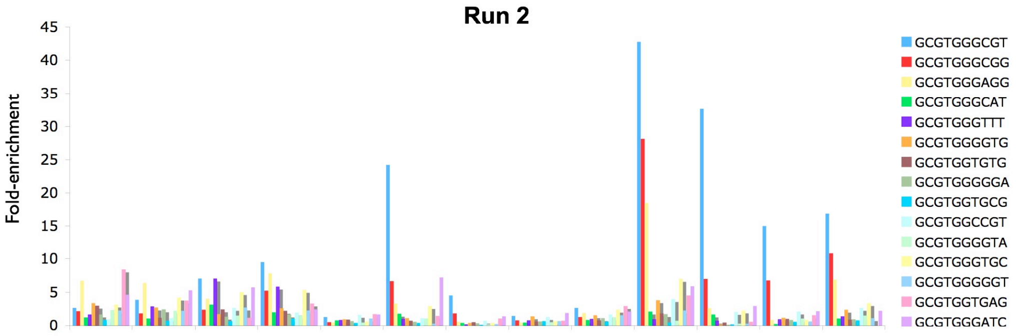
Supplementary Figure S5: Fold-enrichment of 25 motifs. Motifs on the x-axis correspond to motifs shown in Supplementary Figure S5.



Supplementary Figure S6: Fold-enrichment of 15 ten-mers for Zif268.

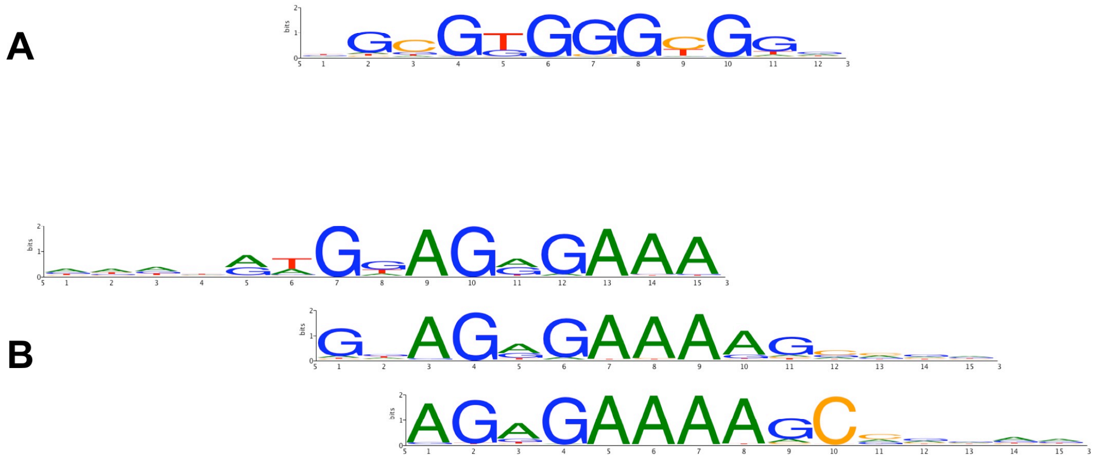


protein	z	z	z	z	z	z	z	z	z
protein (nM)	50	50	50	50	50	5	0.5	0.05	50
salt (mM)	1	10	50	100	500	100	100	100	100
extra									gs

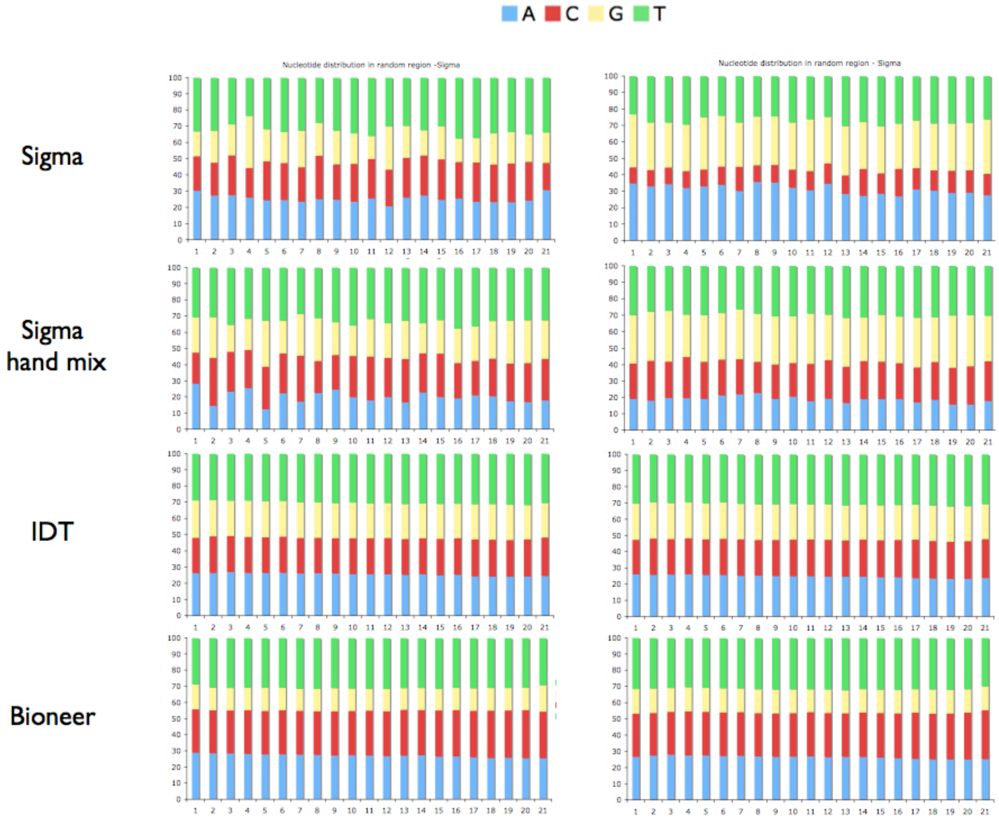


protein	z	z	z	z	z	z	z	z	z	z	z	z	z
protein (nM)	50	50	50	50	50	5	0.5	0.05	50	5	5	5	5
salt (mM)	1	10	50	100	500	100	100	100	100	100	100	200	100
extra									gs	lw	+r		f

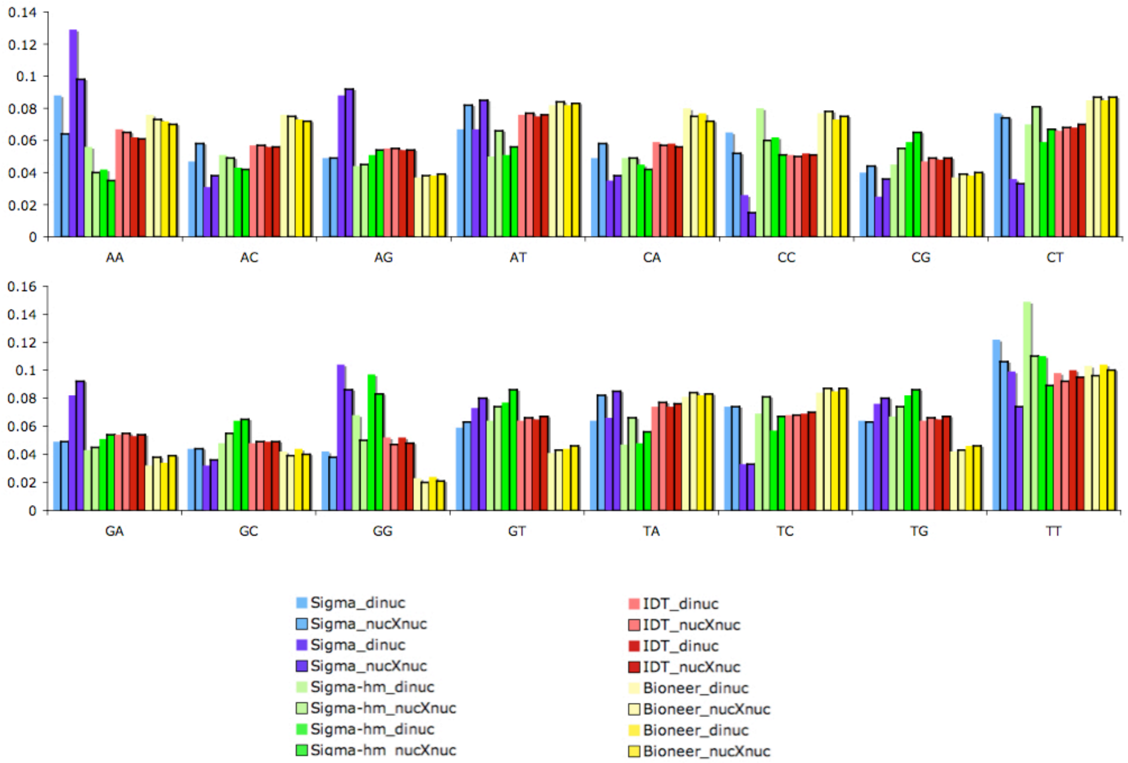
Supplementary Figure S7: MEME results with large windows and multiple motifs. Reads matching the intermediate motifs were used as input to MEME. A) Search performed with Zif268 reads and length = 12. B) Search performed with Aart reads and length = 15.



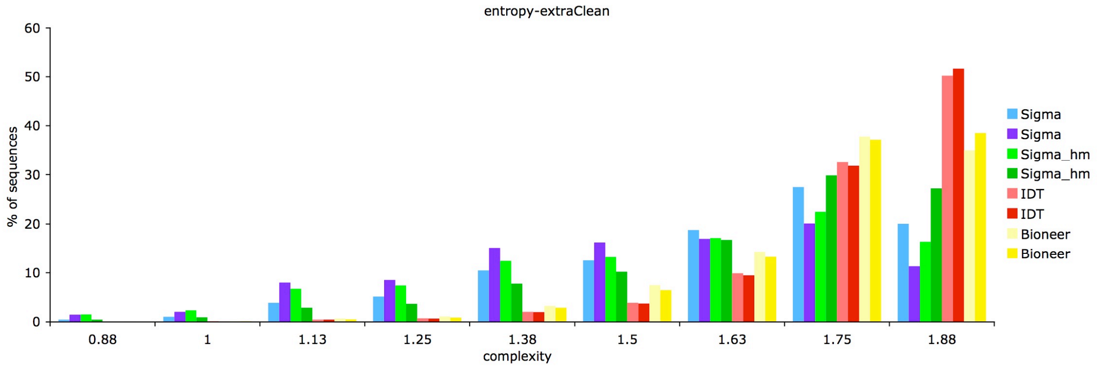
Supplementary Figure S8: Nucleotide composition of the random region.



Supplementary Figure S9: Observed and expected dinucleotide compositions of the random regions. Expected compositions (nucXnuc) follow observed dinucleotide composition (dinuc) in each pair of columns.



Supplementary Figure S10: Complexity (entropy) of the random region. Complexity is calculated as the Shannon information of the nucleotide frequencies. Examples of sequences and their complexities are shown below.



	complexity scale
AAAAAAAAAAAAAAAAAAAAA	0
AAAAAAAAACAAAAAAACAAA	0.5
TAAGAAAAAAAATAAAAAAAA	0.8
AATAAAAAAAAATTAAATAAC	1
ATAACAATAAATATAAGAAAA	1.3
AAGGGAAAGAAGGACGAATAA	1.5
ACCACAAAAATCGAGTAACCA	1.7
GATGACGAATACGTCGTTCTT	2