Supplementary file 1

Time efficiency of motif discovery algorithms used for analyzing ChIP-seq data

Hardware

In order to evaluate the practical usage of the tools for end users, we deliberately ran time efficiency analyses on an intermediate-cost personal laptop: MacBook Intel Core 2 Duo, Processor speed 2GHz, RAM 2Gb, operating system Mac OSX 10.6.7). All programs run about two times faster on a MacBook Pro with 8Gb RAM.

Empirical measurement of time efficiency

The analysis of time efficiency was performed by running each motif discovery algorithm on sequence sets comprising increasing numbers of peaks ranging from 1000 to 10,000,000. All peaks were set to the same width (100 bp) because DREME automatically restricts the analysis to the 100bp around the center of each peak. Random sequences were generated using the RSAT tool *random-seq* with a background model of order 3 trained on mouse non-coding upstream sequences.

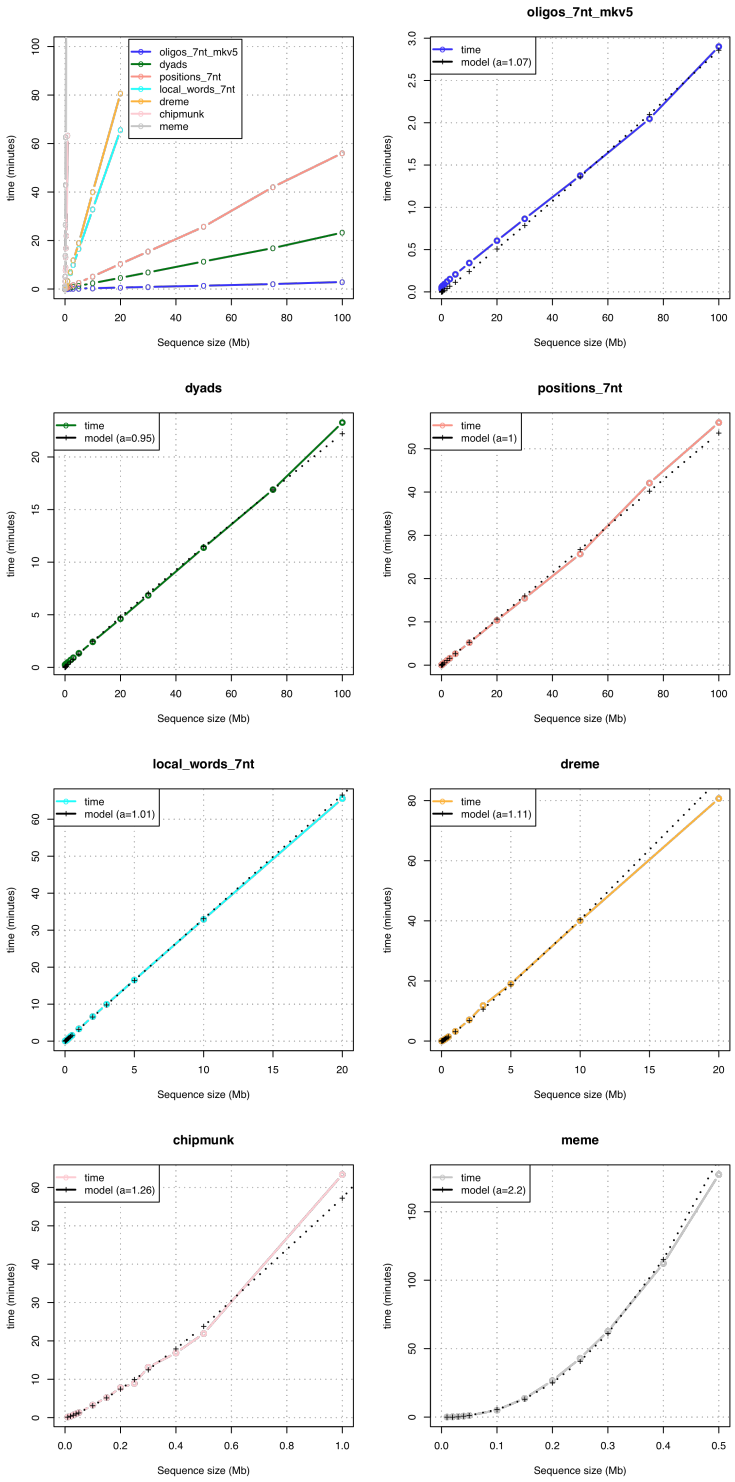
Table S2. **Empirical measurement of time efficiency.** All peaks were set to 100bp for the sake of comparison with dreme. Sequence sizes are in Mb, times in minutes.

Peak nb	Total seq size (Mb)	oligos_7nt_mkv5	local_words_7nt	positions_7nt	dyads	meme	dreme	chipmunk
100	0.01	0.03	0.04	0.03	0.22	0.06	0.01	0.16
200	0.02	0.04	0.06	0.05	0.24	0.20	0.03	0.36
300	0.03	0.05	0.09	0.06	0.24	0.44	0.04	0.69
400	0.04	0.05	0.11	0.07	0.24	0.75	0.05	0.99
500	0.05	0.05	0.13	0.07	0.25	1.16	0.08	1.30
1,000	0.10	0.06	0.23	0.11	0.26	5.09	0.21	3.28
1,500	0.15	0.06	0.36	0.13	0.27	13.71	0.33	5.22
2,000	0.20	0.06	0.56	0.17	0.29	26.61	0.48	7.79
2,500	0.25	0.07	0.75	0.18	0.30	42.94	0.00	8.91
3,000	0.30	0.07	0.91	0.21	0.31	62.65	0.79	13.13
5,000	0.50	0.07	1.58	0.31	0.36	177.17	1.43	21.91
10,000	1	0.09	3.32	0.57	0.48	NA	3.23	63.37
20,000	2	0.12	6.65	1.10	0.70	NA	7.04	NA
30,000	3	0.15	9.97	1.60	0.91	NA	11.82	NA
50,000	5	0.21	16.53	2.63	1.34	NA	19.07	NA
100,000	10	0.34	32.93	5.19	2.42	NA	40.04	NA
200,000	20	0.61	65.58	10.36	4.62	NA	80.68	NA
300,000	30	0.86	NA	15.46	6.85	NA	NA	NA
500,000	50	1.37	NA	25.69	11.37	NA	NA	NA
750,000	75	2.05	NA	42.08	16.90	NA	NA	NA
1,000,000	100	2.90	NA	56.05	23.27	NA	NA	NA

Model fitting

In order to estimate the complexity of each algorithm, we fit a polynomial model onto the observed time response curve. The programs *oligo-analysis*, *dyad-analysis*, *position-analysis* and *dreme* show a linear time complexity (the power is ~1), *chipmunk* has a quasi-linear complexity (power 1.27) and *meme* a more than quadratic complexity (power 2.2).

Figure S1. Comparison of time efficiencies (top right panel) and fitting of a polynomial model to estimate the complexity (all other panels). The power of the polynomial model is displayed in each legend.



Estimation of execution times for representative sequence sizes

Based on the fitted polynomial model, we estimated the time required to treat sequences of 100kb, 1Mb, 10Mb and 100M, respectively. The maximal computing time was limited to about 1 hour for each algorithm, and 3 hours for meme. Estimations are thus interpolations for sizes smaller than the maximal tested size, and extrapolations for sizes exceeding the maximal tested size.

Table S3. Estimation of required time for treating datasets of representative sizes. Expected times were estimated based on the polynomial model fitted on the observed computed times. All times are in minutes.

Algorithm	Max tested size (Mb)	Time for max tested size (min)	power (slope of log linear model)	intercept of log linear model	0.1Mb	1Mb	10Mb	100Mb
oligos_7nt_mkv5	100	2.9	1.07	-3.88	0.00	0.02	0.24	2.86
dyads	100	23.27	0.95	-1.29	0.03	0.28	2.47	22.22
positions_7nt	100	56.05	1.00	-0.64	0.05	0.53	5.32	53.65
local_words_7nt	20	65.58	1.01	1.18	0.32	3.27	33.07	334.86
dreme	20	80.68	1.09	1.18	0.26	3.24	40.12	496.68
chipmunk	1	63.37	1.27	4.06	3.12	58.23	1,086.99	20,290.23
meme	0.5	177.17	2.21	6.78	5.43	881.22	142,998.87	23,204,999.54