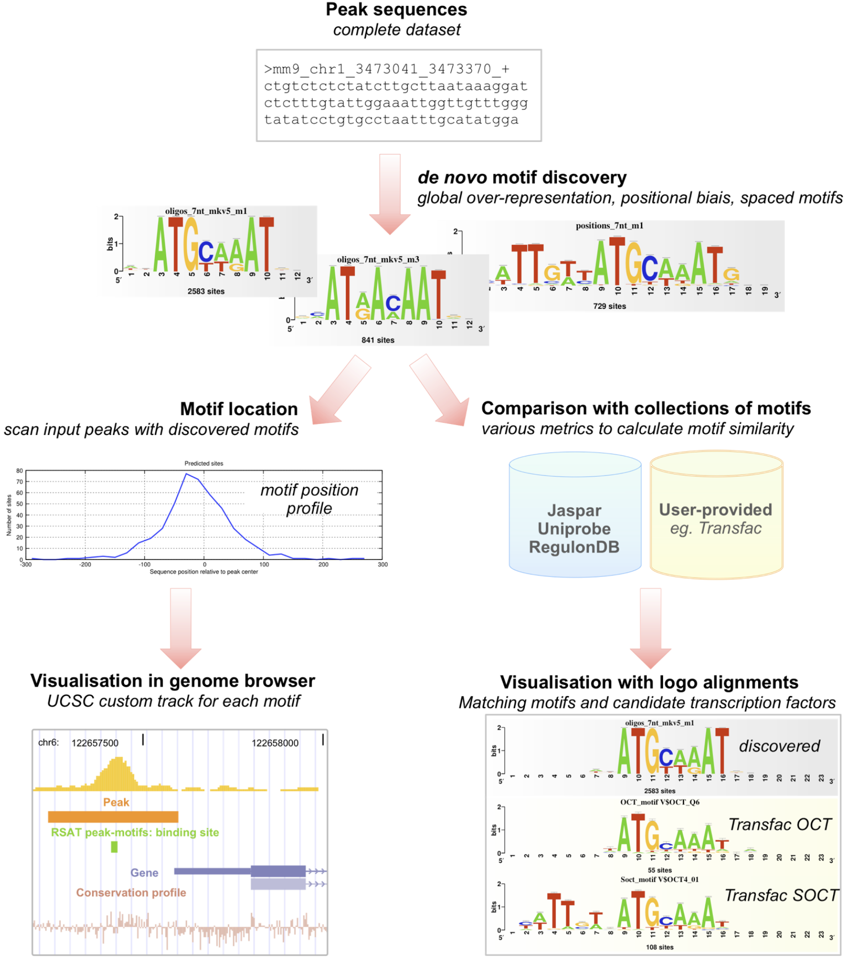
Analyzing ChIP-seq (and related) peaks with RSAT

RSAT peak-motifs : discovering motifs in full-size peak sets

An integrated workflow for analyzing ChIP-seq peaks

- The program ***peak-motifs*** is a work flow combining a series of RSAT tools optimized for discovered motifs in large sequence sets (tens of Mb) resulting from ChIP-seq experiments..
- Multiple pattern discovery algorithms
  - Global over-representation
  - Positional biases
  - Local over-representation
- Discovered motifs are compared with
  - motif databases
  - user-specified reference motifs.
- Prediction of binding sites, which can be uploaded as custom annotation tracks to genome browsers (e.g. UCSC) for visualization.
- Interfaces
  - Stand-alone
  - Web interface
  - Web services (SOAP/WSDL)



Peak sequences
complete dataset

```
>mm9_chr1_3473041_3473370_+
ctgtctctctatcttgcttaataaaggat
ctctttgtattggaaattggttgtttggg
tatatcctgtgcctaatttgcatatgga
```

***de novo* motif discovery**
global over-representation, positional biais, spaced motifs

Motif location
scan input peaks with discovered motifs

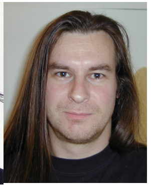
Comparison with collections of motifs
various metrics to calculate motif similarity

Visualisation in genome browser
UCSC custom track for each motif

Visualisation with logo alignments
Matching motifs and candidate transcription factors

Morgane Thomas-Chollier

Matthieu Defrance


Olivier Sand

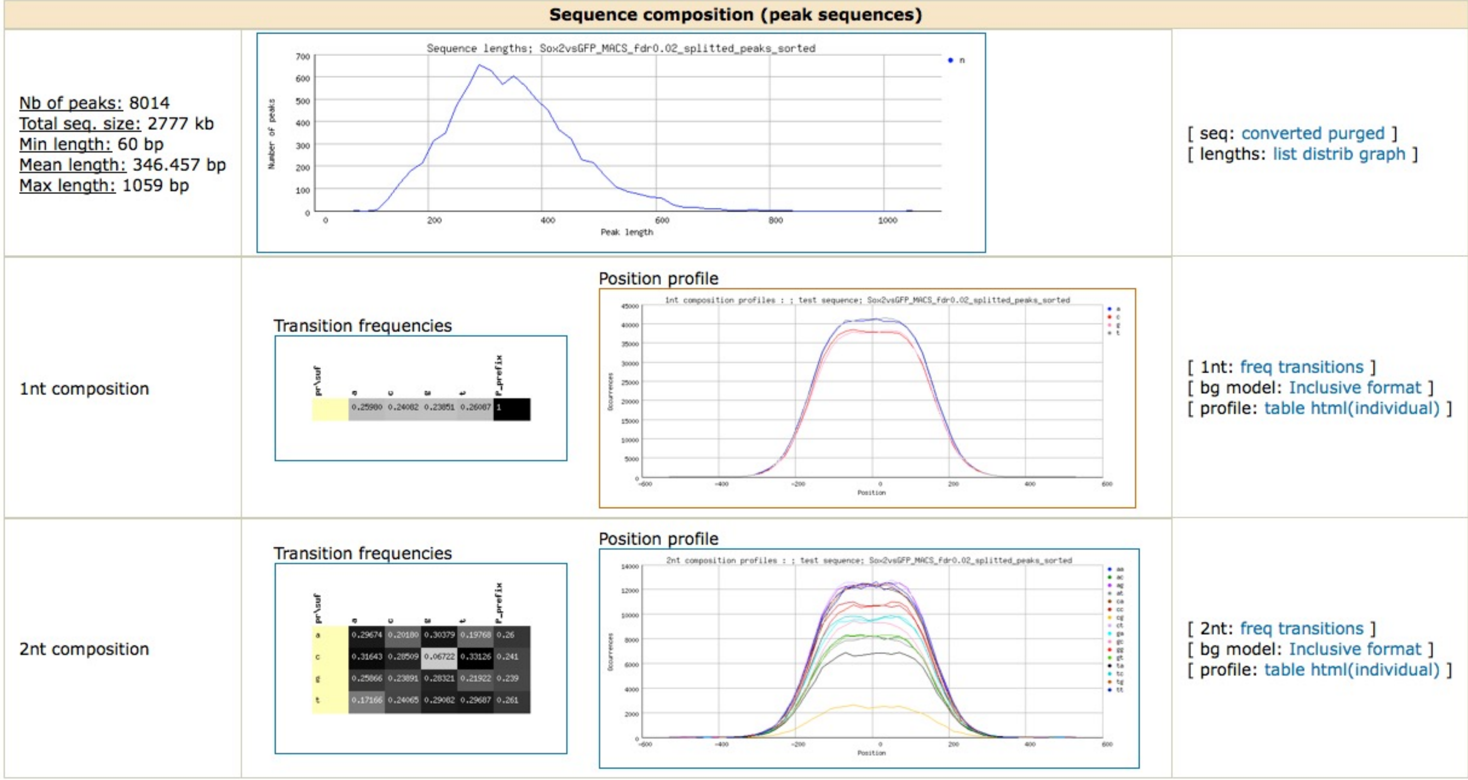
Carl Herrmann

Denis Thieffry

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4): e31.

Composition analysis

- Analysis of the input sequence composition
  - Nucleotide composition + positional distribution
  - Dinucleotide composition reveals dependencies such as CpG islands



Sequence composition (peak sequences)

Nb of peaks: 8014
Total seq. size: 2777 kb
Min length: 60 bp
Mean length: 346.457 bp
Max length: 1059 bp

[seq: converted purged]
[lengths: list distrib graph]

1nt composition

Transition frequencies

pr\suf	a	c	g	t	P_prefix
	0.25980	0.24082	0.23851	0.26087	1

Position profile

[1nt: freq transitions]
[bg model: Inclusive format]
[profile: table html(individual)]

2nt composition

Transition frequencies

pr\suf	a	c	g	t	P_prefix
a	0.29674	0.20180	0.30379	0.19768	0.26
c	0.31643	0.28509	0.06722	0.33126	0.241
g	0.25866	0.23891	0.28321	0.21922	0.239
t	0.17166	0.24065	0.29082	0.29687	0.261

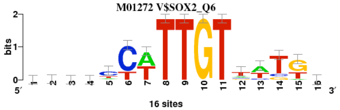
Position profile

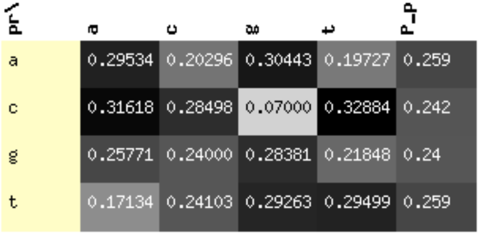
[2nt: freq transitions]
[bg model: Inclusive format]
[profile: table html(individual)]

Composition analysis results

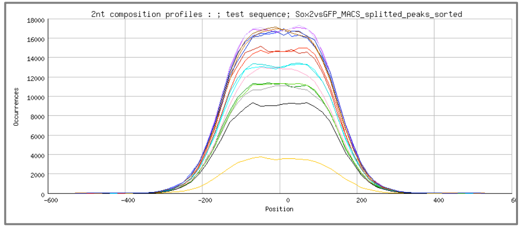
- The composition analysis reveals differences between data sets.
  - Sox2 and Ctcf peaks: clear avoidance of CpG dinucleotides.
  - n-Myc peaks appear as CpG island (the avoidance of CpG is relaxed).
  - The center of Ctcf peaks shows a strong depletion in AA, TT, AT and TA.

Sox2

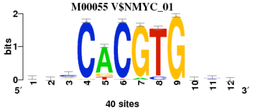


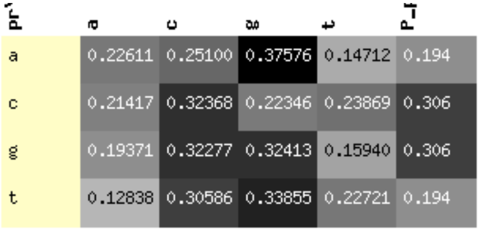


pr\	a	c	g	t	P_P
a	0.29534	0.20296	0.30443	0.19727	0.259
c	0.31618	0.28498	0.07000	0.32884	0.242
g	0.25771	0.24000	0.28381	0.21848	0.24
t	0.17134	0.24103	0.29263	0.29499	0.259

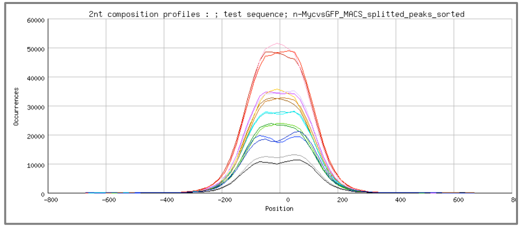


n-Myc

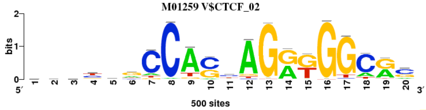


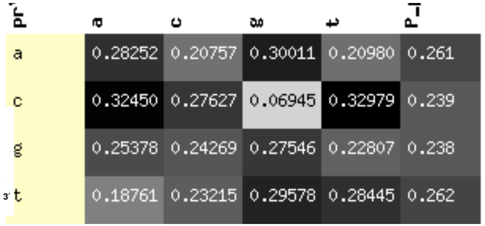


pr\	a	c	g	t	P_f
a	0.22611	0.25100	0.37576	0.14712	0.194
c	0.21417	0.32368	0.22346	0.23869	0.306
g	0.19371	0.32277	0.32413	0.15940	0.306
t	0.12838	0.30586	0.33855	0.22721	0.194

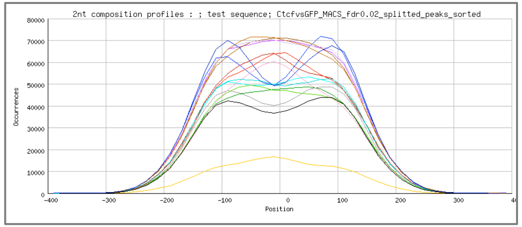


Ctcf





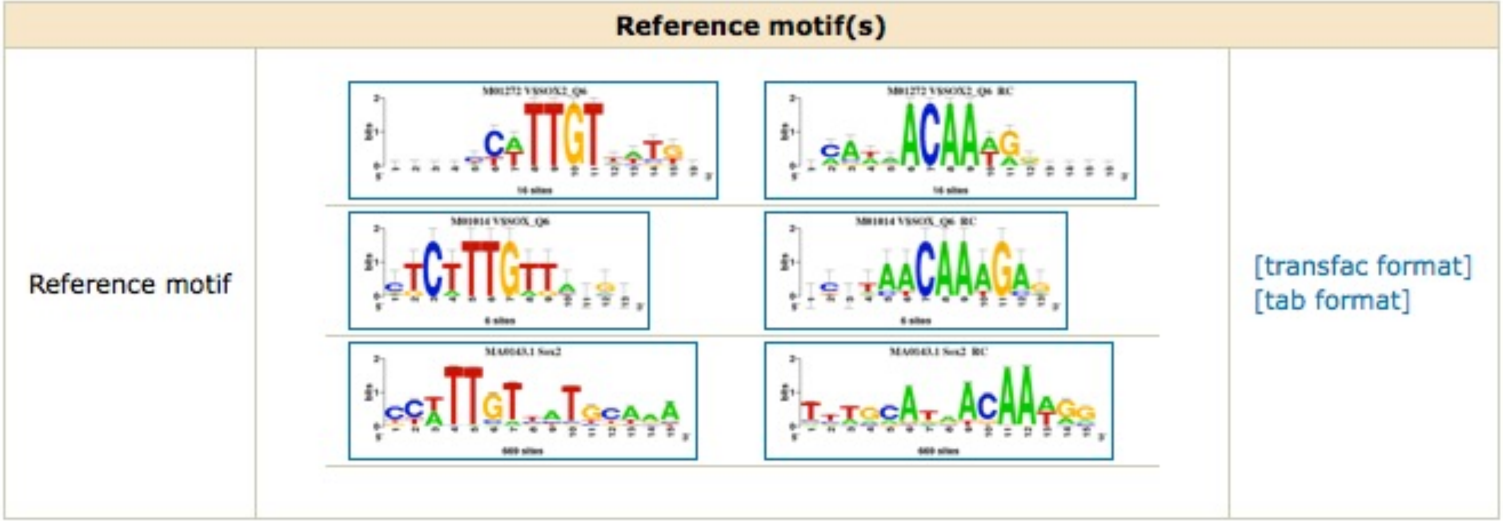
pr\	a	c	g	t	P_l
a	0.28252	0.20757	0.30011	0.20980	0.261
c	0.32450	0.27627	0.06945	0.32979	0.239
g	0.25378	0.24269	0.27546	0.22807	0.238
t	0.18761	0.23215	0.29578	0.28445	0.262



● aa
● ac
● ag
● at
● ca
● cc
● cg
● ct
● ga
● gc
● gg
● gt
● ta
● tc
● tg
● tt

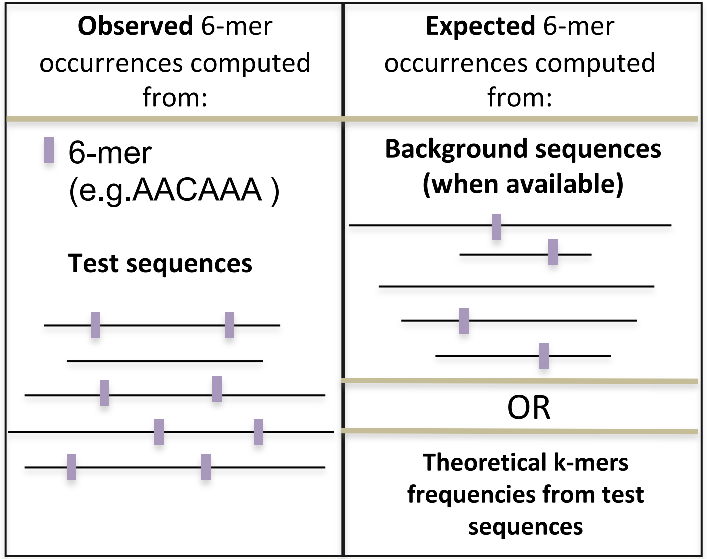
User-specified reference motifs (the “expected” answer)

- One or several reference motifs can be defined.
- Reference motifs are the ones which are expected to be found in the dataset.
  - More precisely, if those motifs are not reported, it is considered as a failure.
- Choice of reference motifs is somewhat tricky.
  - Example: Sox2 peaks
    - 2 slightly different matrices are annotated in TRANSFAC for Sox2
    - The 3rd matrix reflects the composite Sox/Oct motif (SOCT).
    - This motif was obtained by the TRANSFAC team using a motif discovery algorithm on Chen data set -> not properly speaking a “golden reference” for evaluating motif discovery accuracy.



Reference motif(s)		
Reference motif		[transfac format] [tab format]

Detection of global over- or under-representation



Observed 6-mer occurrences computed from:

6-mer (e.g.AACAAA)

Test sequences

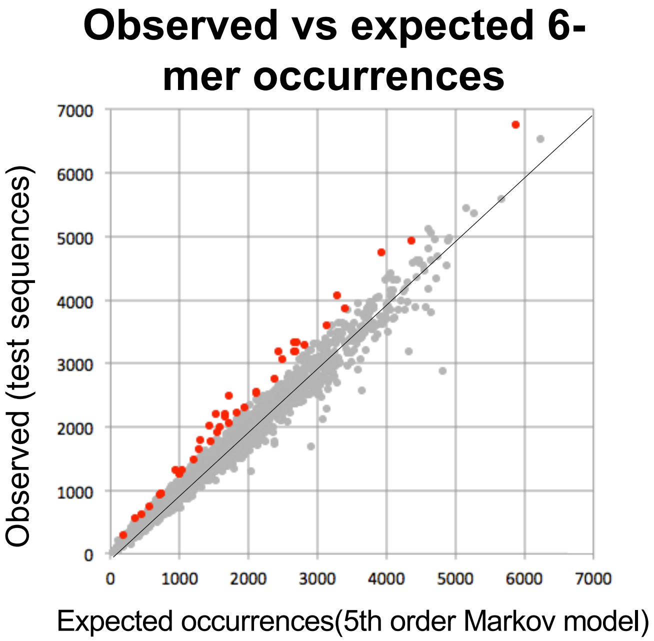
Expected 6-mer occurrences computed from:

Background sequences (when available)

OR

Theoretical k-mers frequencies from test sequences

➔ computation of p-value (binomial) and E-value (multi-testing correction)



Observed vs expected 6-mer occurrences

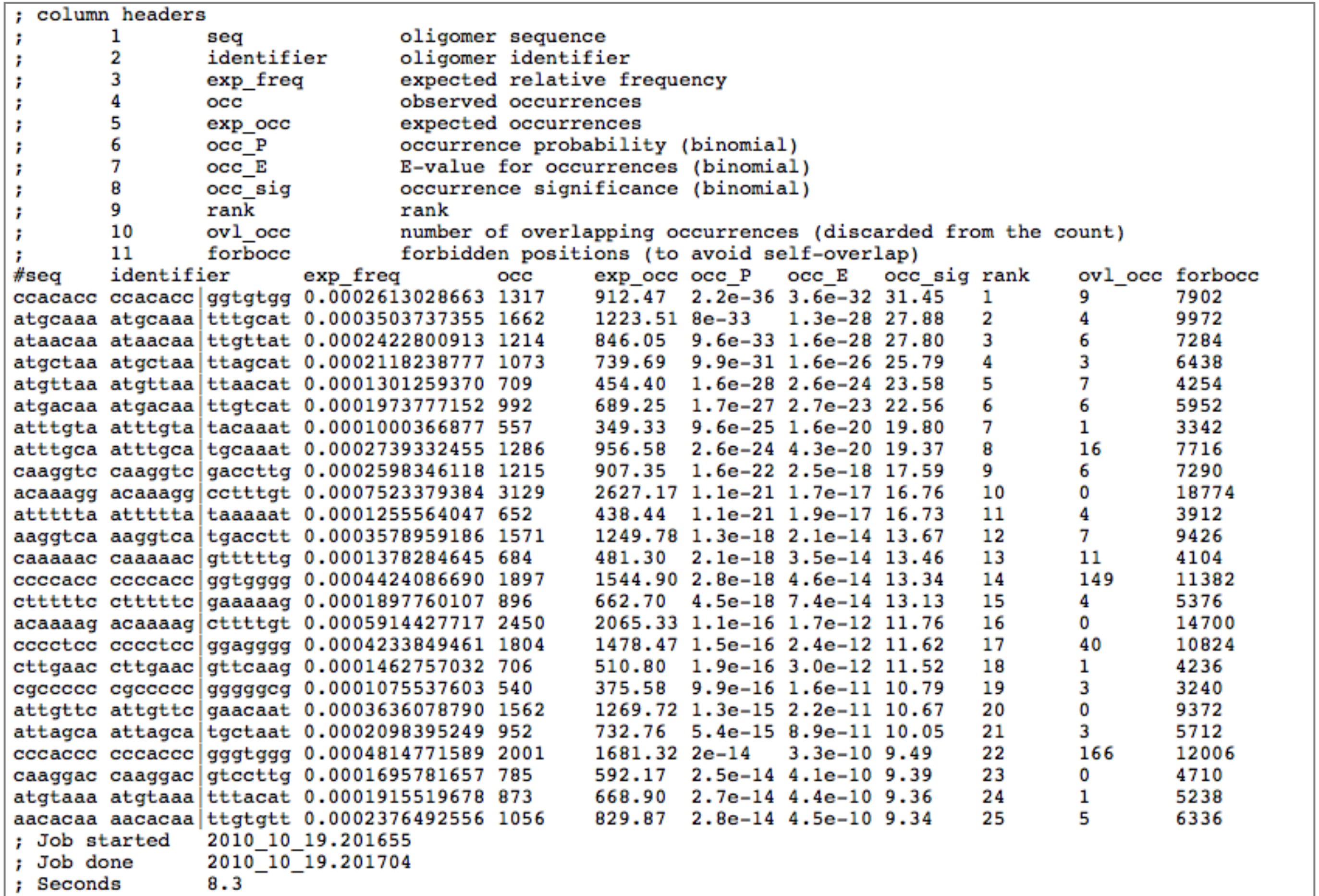
oligo-analysis and dyad-analysis

1. van Helden, J., Andre, B. and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281, 827-42.
2. van Helden, J., del Olmo, M. and Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28, 1000-10.
3. van Helden, J., Rios, A. F. and Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28, 1808-18.

Peak-motifs

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. Nat Protoc 7(8): 1551-1568.

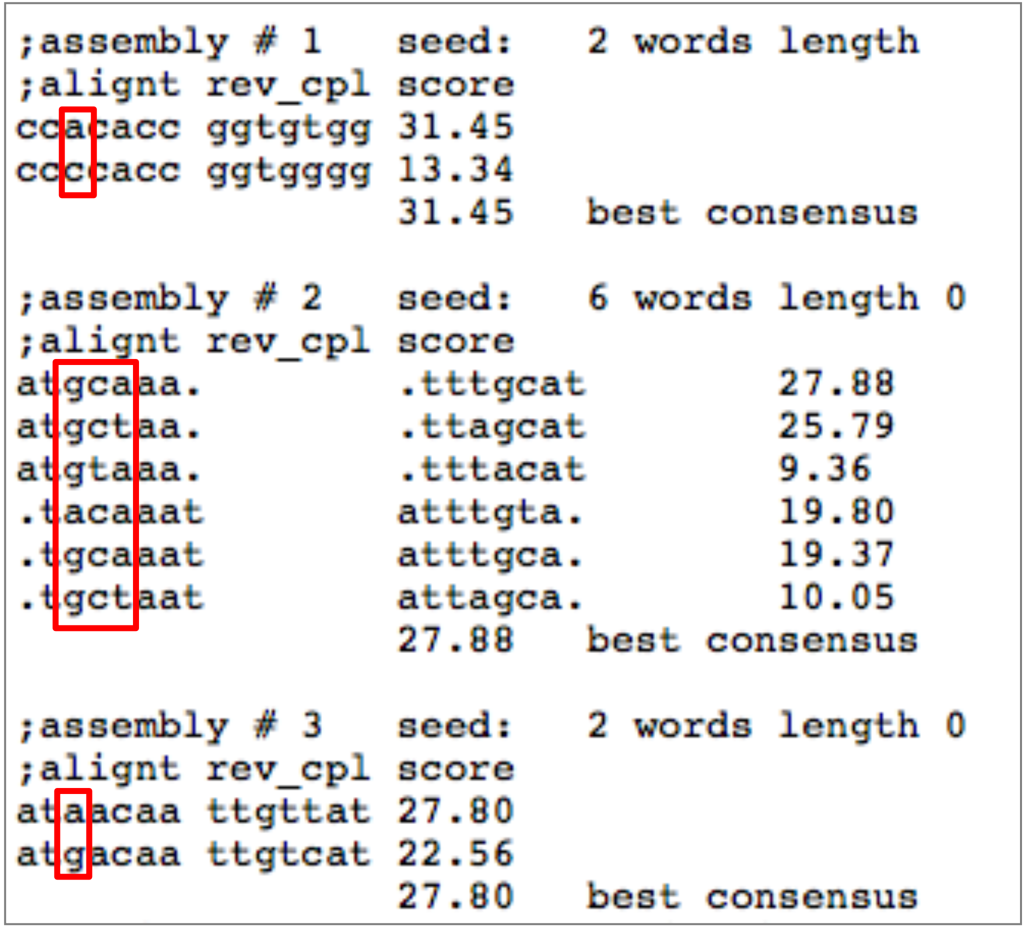
Primary result: a list of over-represented words



```
; column headers
;       1       seq             oligomer sequence
;       2       identifier      oligomer identifier
;       3       exp_freq        expected relative frequency
;       4       occ             observed occurrences
;       5       exp_occ         expected occurrences
;       6       occ_P           occurrence probability (binomial)
;       7       occ_E           E-value for occurrences (binomial)
;       8       occ_sig         occurrence significance (binomial)
;       9       rank            rank
;       10      ovl_occ         number of overlapping occurrences (discarded from the count)
;       11      forbocc         forbidden positions (to avoid self-overlap)
#seq    identifier      exp_freq        occ     exp_occ occ_P   occ_E   occ_sig rank    ovl_occ forbocc
ccacacc ccacacc|ggtgtgg 0.0002613028663 1317    912.47  2.2e-36 3.6e-32 31.45   1       9       7902
atgcaaa atgcaaa|tttgcat 0.0003503737355 1662    1223.51 8e-33   1.3e-28 27.88   2       4       9972
ataacaa ataacaa|ttgttat 0.0002422800913 1214    846.05  9.6e-33 1.6e-28 27.80   3       6       7284
atgctaa atgctaa|ttagcat 0.0002118238777 1073    739.69  9.9e-31 1.6e-26 25.79   4       3       6438
atgttaa atgttaa|ttaacat 0.0001301259370 709     454.40  1.6e-28 2.6e-24 23.58   5       7       4254
atgacaa atgacaa|ttgtcat 0.0001973777152 992     689.25  1.7e-27 2.7e-23 22.56   6       6       5952
atttgta atttgta|tacaaat 0.0001000366877 557     349.33  9.6e-25 1.6e-20 19.80   7       1       3342
atttgca atttgca|tgcaaat 0.0002739332455 1286    956.58  2.6e-24 4.3e-20 19.37   8       16      7716
caaggtc caaggtc|gaccttg 0.0002598346118 1215    907.35  1.6e-22 2.5e-18 17.59   9       6       7290
acaaagg acaaagg|cctttgt 0.0007523379384 3129    2627.17 1.1e-21 1.7e-17 16.76   10      0       18774
atttta atttta|taaaaat 0.0001255564047 652     438.44  1.1e-21 1.9e-17 16.73   11      4       3912
aaggtca aaggtca|tgacctt 0.0003578959186 1571    1249.78 1.3e-18 2.1e-14 13.67   12      7       9426
caaaaac caaaaac|gtttttg 0.0001378284645 684     481.30  2.1e-18 3.5e-14 13.46   13      11      4104
ccccacc ccccacc|ggtgggg 0.0004424086690 1897    1544.90 2.8e-18 4.6e-14 13.34   14      149     11382
ctttttc ctttttc|gaaaaag 0.0001897760107 896     662.70  4.5e-18 7.4e-14 13.13   15      4       5376
acaaaag acaaaag|cttttgt 0.0005914427717 2450    2065.33 1.1e-16 1.7e-12 11.76   16      0       14700
cccctcc cccctcc|ggagggg 0.0004233849461 1804    1478.47 1.5e-16 2.4e-12 11.62   17      40      10824
cttgaac cttgaac|gttcaag 0.0001462757032 706     510.80  1.9e-16 3.0e-12 11.52   18      1       4236
cgccccc cgccccc|gggggcg 0.0001075537603 540     375.58  9.9e-16 1.6e-11 10.79   19      3       3240
attgttc attgttc|gaacaat 0.0003636078790 1562    1269.72 1.3e-15 2.2e-11 10.67   20      0       9372
attagca attagca|tgctaat 0.0002098395249 952     732.76  5.4e-15 8.9e-11 10.05   21      3       5712
cccaccc cccaccc|gggtggg 0.0004814771589 2001    1681.32 2e-14   3.3e-10 9.49    22      166     12006
caaggac caaggac|gtccttg 0.0001695781657 785     592.17  2.5e-14 4.1e-10 9.39    23      0       4710
atgtaaa atgtaaa|tttacat 0.0001915519678 873     668.90  2.7e-14 4.4e-10 9.36    24      1       5238
aacacaa aacacaa|ttgtgtt 0.0002376492556 1056    829.87  2.8e-14 4.5e-10 9.34    25      5       6336
; Job started   2010_10_19.201655
; Job done      2010_10_19.201704
; Seconds       8.3
```

Over-represented words reveal motif variability

- The list of over-represented words generally contain groups of mutually overlapping words.
- Those groups can be aligned using the program *pattern-assembly*
- Assembled words reveal
  - larger motifs than the initial word length
  - positions with variable residues
- Word assemblies can be used to build a matrix.
  - Assembled words are used as seed to scan input sequences for sites.
  - A new matrix is build from the collected sites.

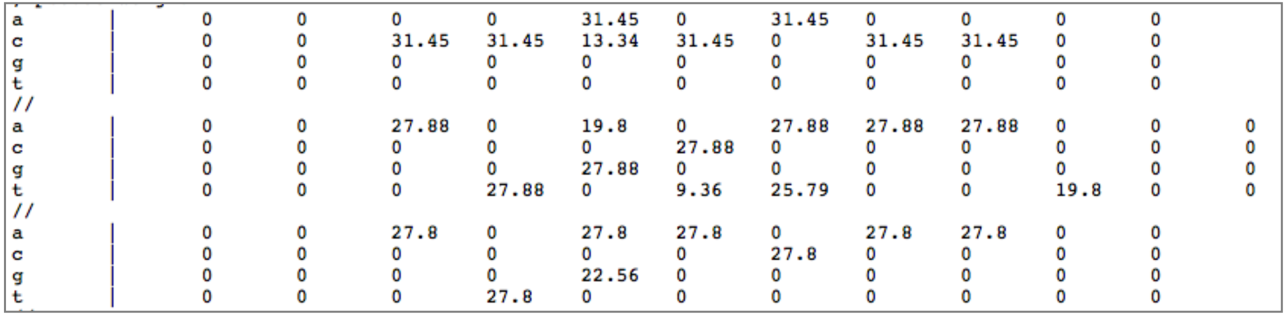


```
;assembly # 1   seed:   2 words length
;alignt rev_cpl score
ccacacc ggtgtgg 31.45
ccccacc ggtgggg 13.34
                31.45   best consensus

;assembly # 2   seed:   6 words length 0
;alignt rev_cpl score
atgcaaa.        .tttgcat        27.88
atgctaa.        .ttagcat        25.79
atgtaaa.        .tttacat        9.36
.tacaaat        atttgta.        19.80
.tgcaaat        atttgca.        19.37
.tgctaat        attagca.        10.05
                27.88   best consensus

;assembly # 3   seed:   2 words length 0
;alignt rev_cpl score
ataacaa ttgttat 27.80
atgacaa ttgtcat 22.56
                27.80   best consensus
```

Collecting a matrix from assembled words

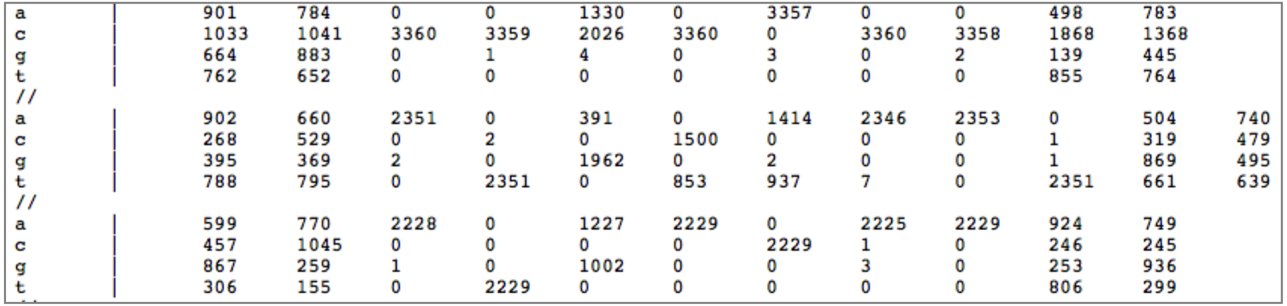
Significance matrix

- The significance matrix can be used as “seed” to scan the input sequences and collect binding sites.
- Those sites are in turn used to build a final matrix.

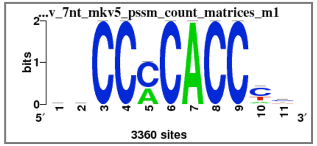


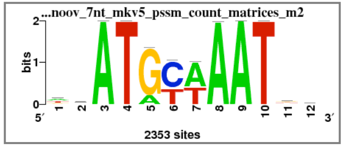
```
a    |    0      0      0      0      31.45  0      31.45  0      0      0      0
c    |    0      0      31.45  31.45  13.34  31.45  0      31.45  31.45  0      0
g    |    0      0      0      0      0      0      0      0      0      0      0
t    |    0      0      0      0      0      0      0      0      0      0      0
//
a    |    0      0      27.88  0      19.8   0      27.88  27.88  27.88  0      0      0
c    |    0      0      0      0      0      27.88  0      0      0      0      0      0
g    |    0      0      0      0      27.88  0      0      0      0      0      0      0
t    |    0      0      0      27.88  0      9.36   25.79  0      0      19.8   0      0
//
a    |    0      0      27.8   0      27.8   27.8   0      27.8   27.8   0      0
c    |    0      0      0      0      0      0      27.8   0      0      0      0
g    |    0      0      0      0      22.56  0      0      0      0      0      0
t    |    0      0      0      27.8   0      0      0      0      0      0      0
```

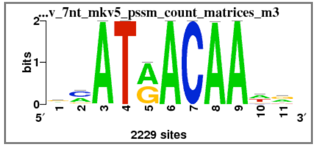
Final matrix



```
a    |    901    784    0      0      1330   0      3357   0      0      498    783
c    |    1033   1041   3360   3359   2026   3360   0      3360   3358   1868   1368
g    |    664    883    0      1      4      0      3      0      2      139    445
t    |    762    652    0      0      0      0      0      0      0      855    764
//
a    |    902    660    2351   0      391    0      1414   2346   2353   0      504    740
c    |    268    529    0      2      0      1500   0      0      0      1      319    479
g    |    395    369    2      0      1962   0      2      0      0      1      869    495
t    |    788    795    0      2351   0      853    937    7      0      2351   661    639
//
a    |    599    770    2228   0      1227   2229   0      2225   2229   924    749
c    |    457    1045   0      0      0      0      2229   1      0      246    245
g    |    867    259    1      0      1002   0      0      3      0      253    936
t    |    306    155    0      2229   0      0      0      0      0      806    299
```

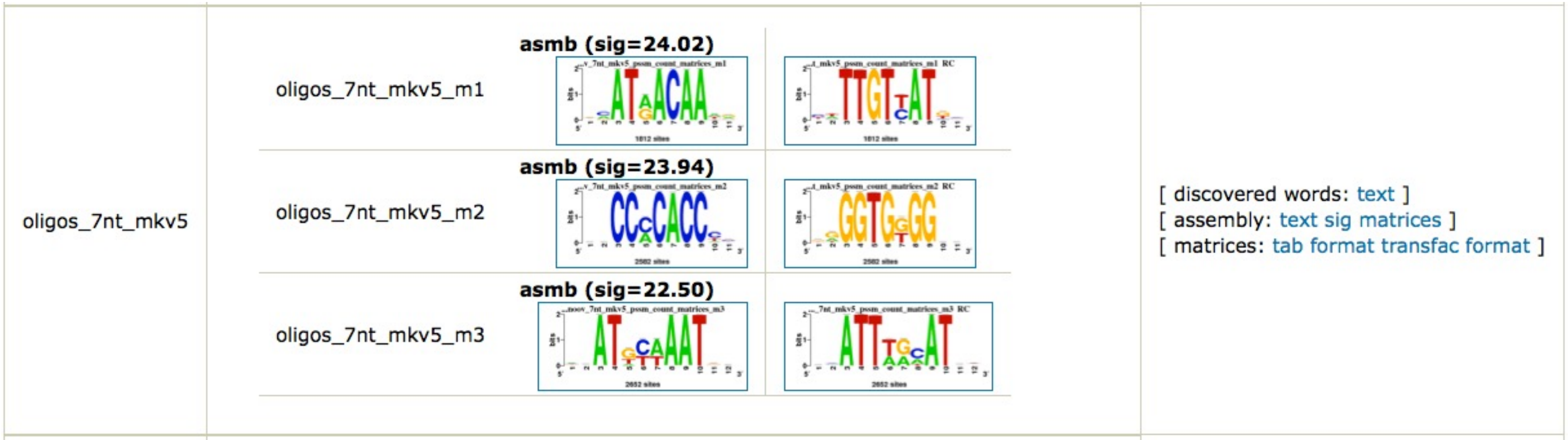






Motifs reported with oligo-analysis (Sox2 peaks from Chen, 2008)

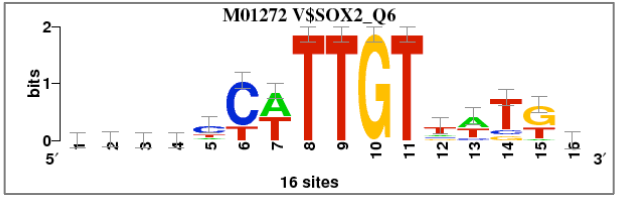
- ***oligo-analysis*** detects over-represented k-mers, as compared to some background model.
- For length k, we use the most stringent Markov chain model (m = k – 2).
- The program detects the Sox2 and Oct4 motifs.
- It also returns a Klf-like motif



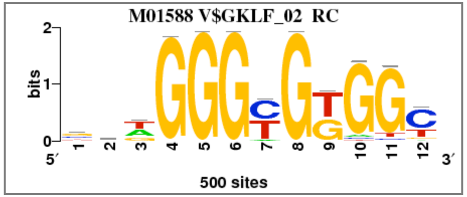
oligos_7nt_mkv5	oligos_7nt_mkv5_m1	asmb (sig=24.02)		[discovered words: text] [assembly: text sig matrices] [matrices: tab format transfac format]
	oligos_7nt_mkv5_m2	asmb (sig=23.94)		
	oligos_7nt_mkv5_m3	asmb (sig=22.50)		

...v_7nt_mkv5_pssm_count_matrices_m1
ATAACAA
1812 sites
...t_mkv5_pssm_count_matrices_m1 RC
TTGTTAT
...v_7nt_mkv5_pssm_count_matrices_m2
CCACACC
2582 sites
...t_mkv5_pssm_count_matrices_m2 RC
GGTGGG
...noov_7nt_mkv5_pssm_count_matrices_m3
ATGCAAAT
2652 sites
..._7nt_mkv5_pssm_count_matrices_m3 RC
ATTTGCAT

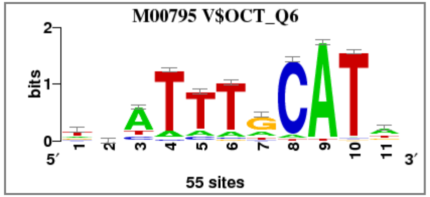
Sox2 (TRANSFAC, built from individual sites)



KLF (TRANSFAC built from Chen Klf4 set)



OCT (TRANSFAC, various OCT factors)



Detecting heterogeneous repartition along sequences

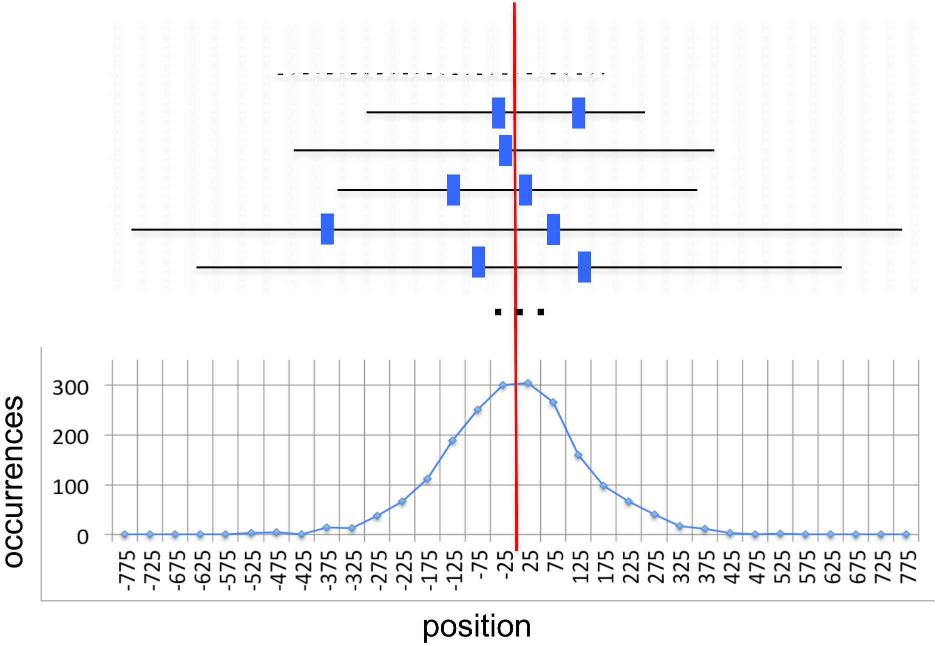
Observed occurrences per window

Expected occurrences per window according to an homogeneous model

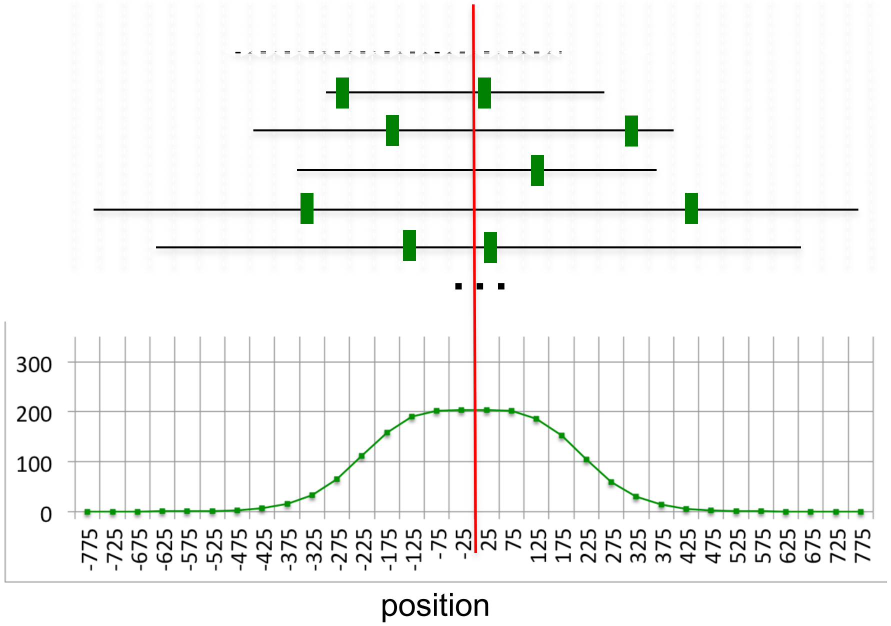
occurrences

🟦🟩 7-mer (e.g.AACAAAG)

Windows 50 nt



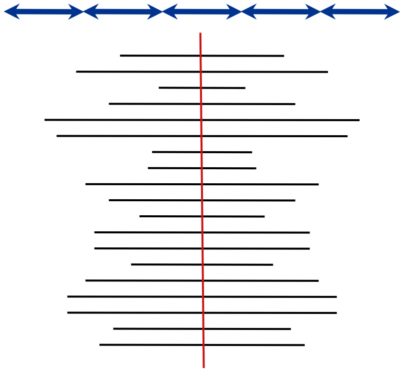
Windows 50 nt

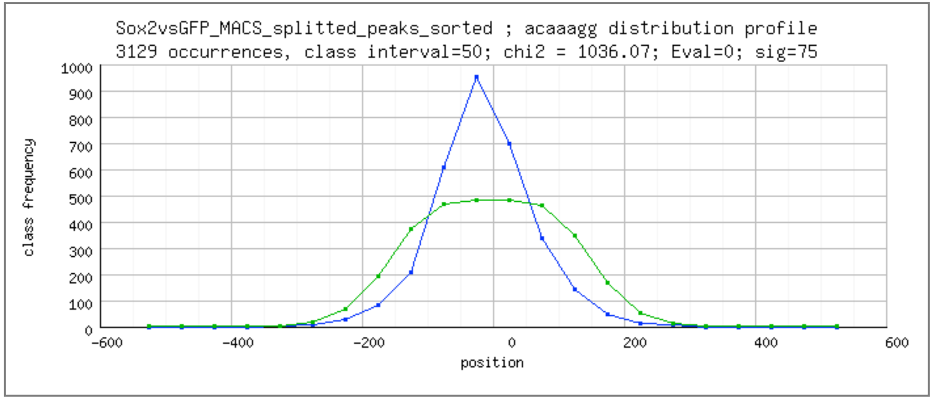


Drawing by Elodie Darbo.

Detecting k-mers with biased positional distribution

- ***position-analysis*** (van Helden et al., 2000) detects k-mers having a heterogeneous distribution of occurrences across input sequences.
- Principle: for each k-mer
  - Compute the number of occurrences in non-overlapping windows starting from a reference point (sequence start, center or end).
  - Compute the expected occurrences in each window according to a homogeneous distribution model.
  - Compute the difference between the observed and expected positional distribution (chi2 test for goodness of fit).
- Example: Sox2 peaks from Chen, 2008
  - 10,929 peaks of size between 60 and 1,059 bp
  - Length : k=7
  - Reference position: the center of each peak.
  - The most significant k-mer is ACAAAGG, which corresponds to the Sox2 consensus.




Sox2vsGFP_MACS_splitted_peaks_sorted ; acaaagg distribution profile
3129 occurrences, class interval=50; chi2 = 1036.07; Eval=0; sig=75
class frequency
1000
900
800
700
600
500
400
300
200
100
0
-600
-400
-200
0
200
400
600
position

Green: expected occurrences

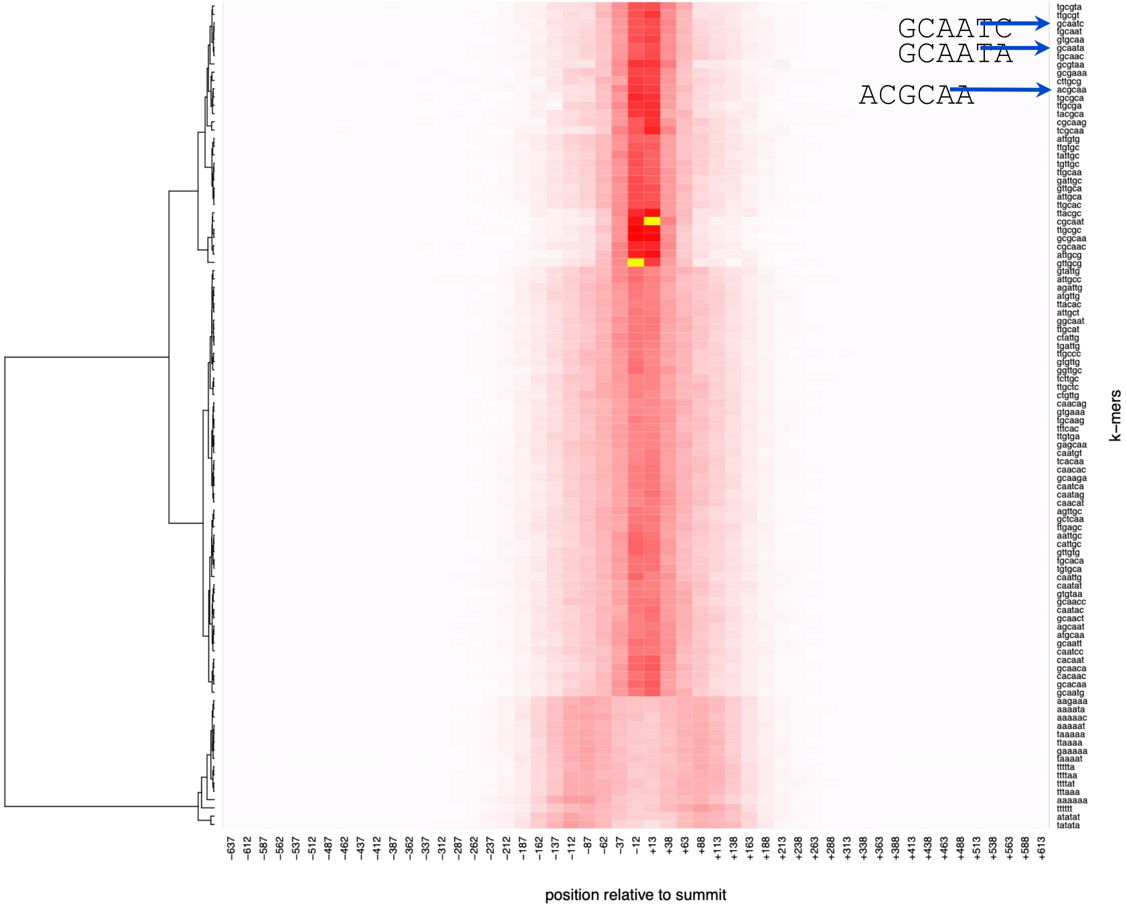
- Note: the expectation decreases with the distance to peak center because peaks have variable lengths.

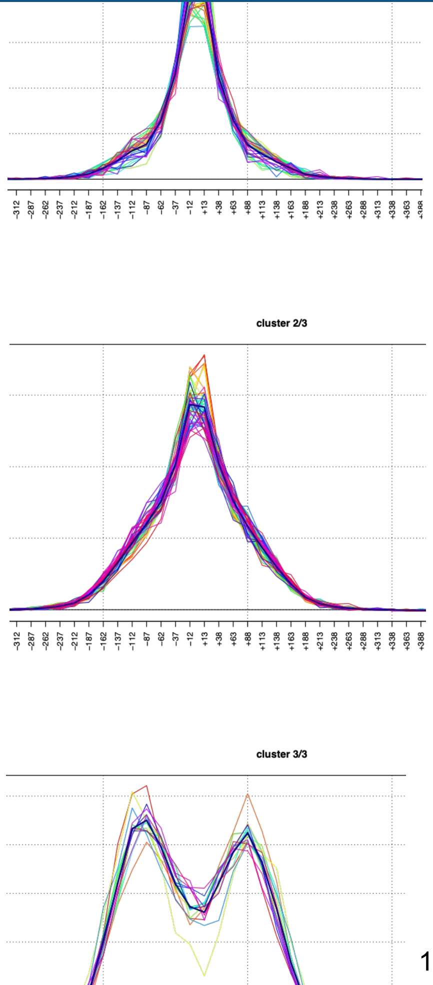
Blue: observed occurrences

- The k-mer ACAAGG is concentrated the center the ChIP-seq peak regions.

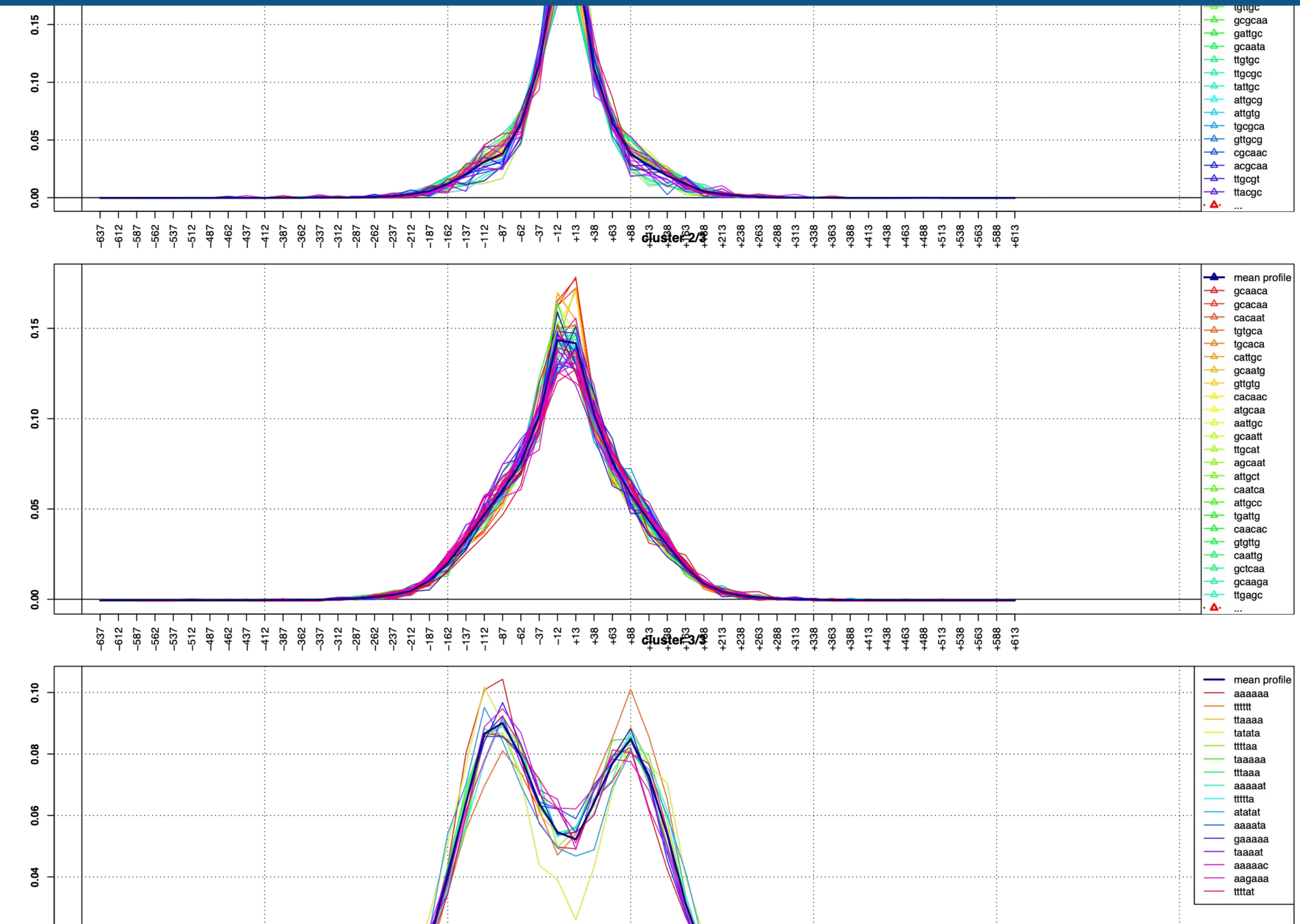
van Helden, J., del Olmo, M. and Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28, 1000-10.

Clustering k-mers by positional density profiles



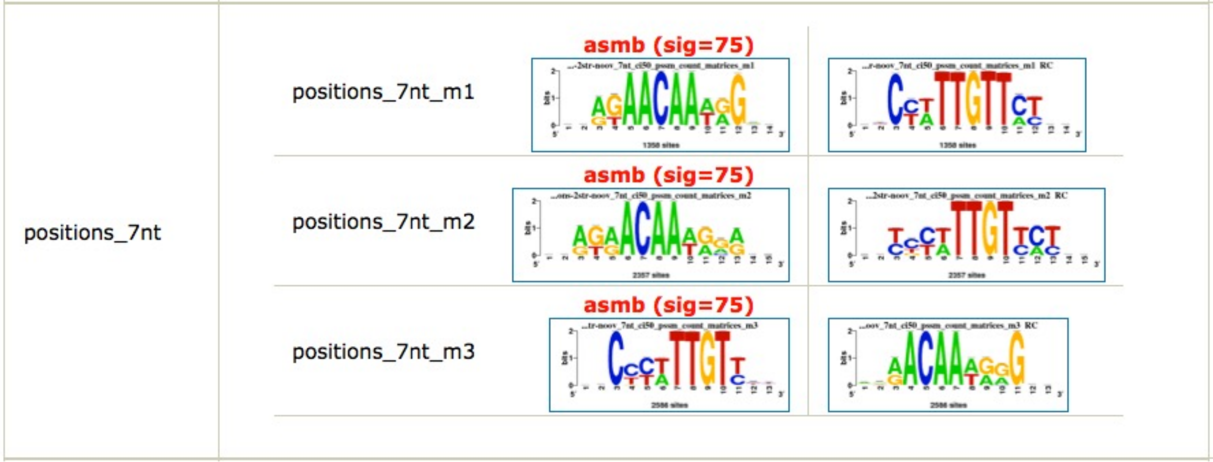


Position profile clustering



Motifs with position biases in Sox2 peaks from Chen, 2008

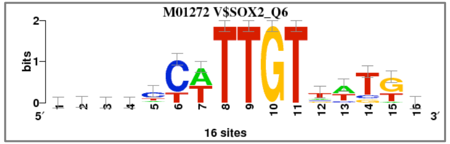
- position-analysis
  - detects the Sox2 motif in Sox2 peaks (redundant motifs are found by different assemblies of oligonucleotides).
  - The motifs of partner TFs (Oct4, Klf4) are not detected.



positions_7nt	positions_7nt_m1	asmb (sig=75)	
	positions_7nt_m2	asmb (sig=75)	
	positions_7nt_m3	asmb (sig=75)	

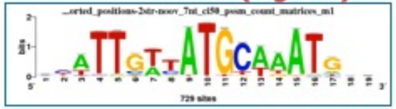
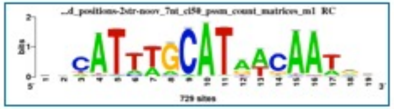
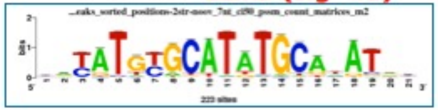
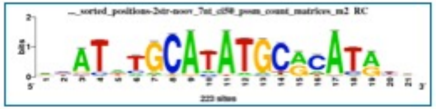
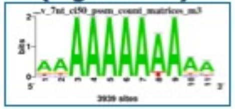
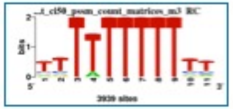
...-2str-noov_7nt_ci50_pssm_count_matrices_m1
1358 sites
...r-noov_7nt_ci50_pssm_count_matrices_m1 RC
1358 sites
...ots-2str-noov_7nt_ci50_pssm_count_matrices_m2
2357 sites
...2str-noov_7nt_ci50_pssm_count_matrices_m2 RC
2357 sites
...tr-noov_7nt_ci50_pssm_count_matrices_m3
2586 sites
...oov_7nt_ci50_pssm_count_matrices_m3 RC
2586 sites
bits

Sox2 (TRANSFAC, built from individual sites)

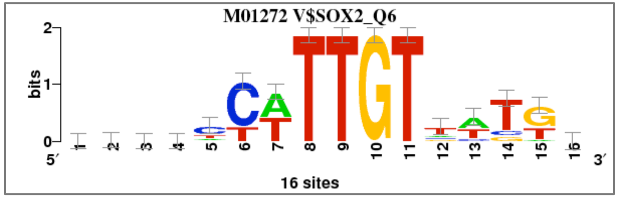


Motifs with position biases in Oct4 peaks from Chen, 2008

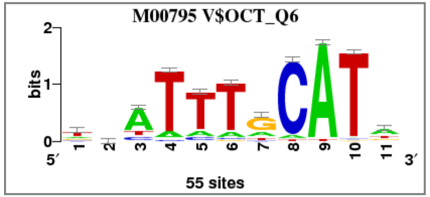
- position-analysis
  - detects the hybrid Sox/Oct motif in Oct4 peaks

positions_7nt	positions_7nt_m1	asmb (sig=75)  	
	positions_7nt_m2	asmb (sig=75)  	
	positions_7nt_m3	asmb (sig=38.13)  	

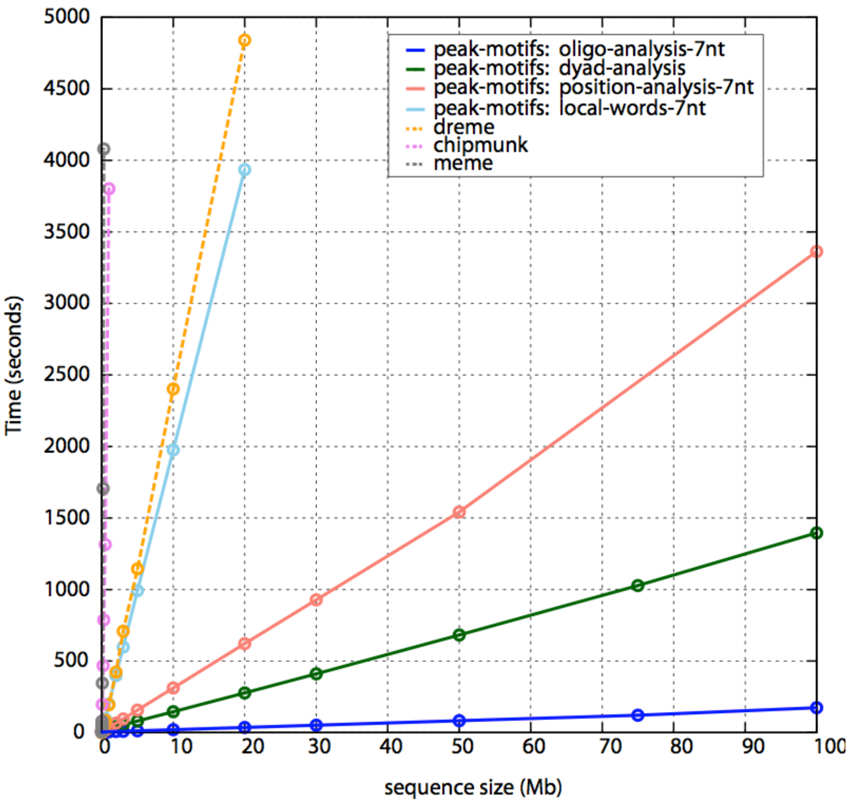
Sox2 (TRANSFAC, built from individual sites)



OCT (TRANSFAC, various OCT factors)



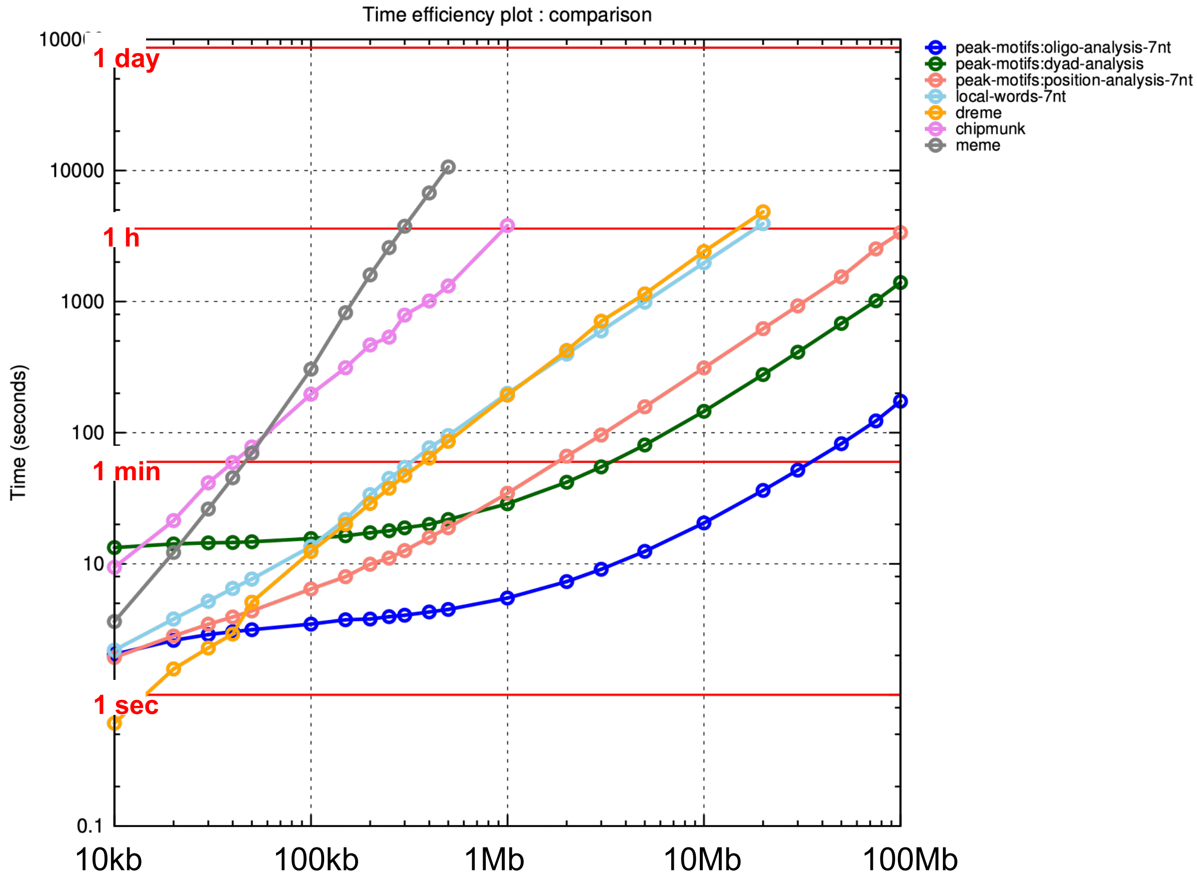
Time efficiency



- String-based approaches
  - The processing time increases linearly with sequence size.
  - The memory is principally affected by the number of patterns (oligo size) -> large sequences can be treated with moderate RAM.
- MEME
  - Processing time is quadratic.
- On a medium-priced laptop (MacBook, 2Gb RAM), the biggest files (100Mb) is treated in
  - 3 minuteswith oligo-analysis;
  - 25 minutes with dyad-analysis;
  - <1 hour with position-analysis.
  - 44 years with meme (polynomial extrapolation)

- Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4): e31.

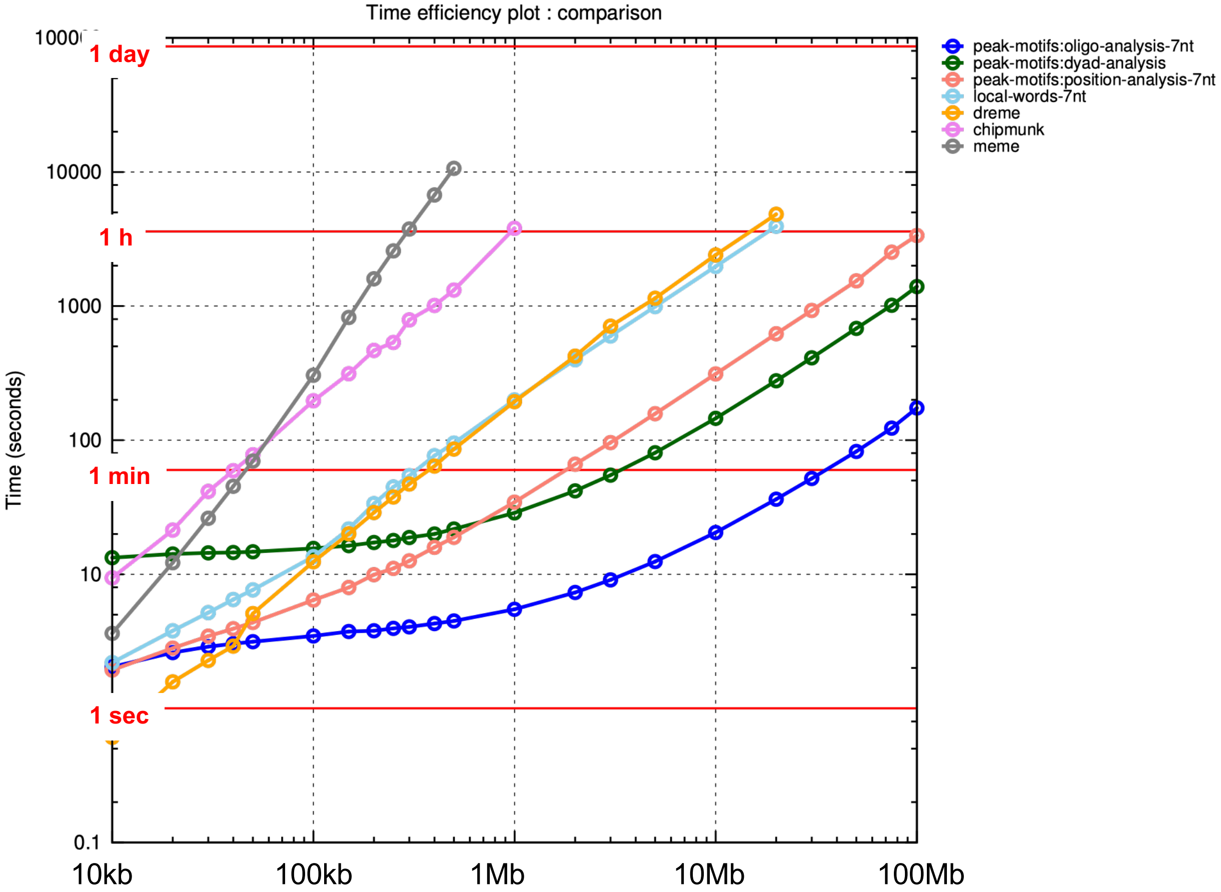
Time efficiency



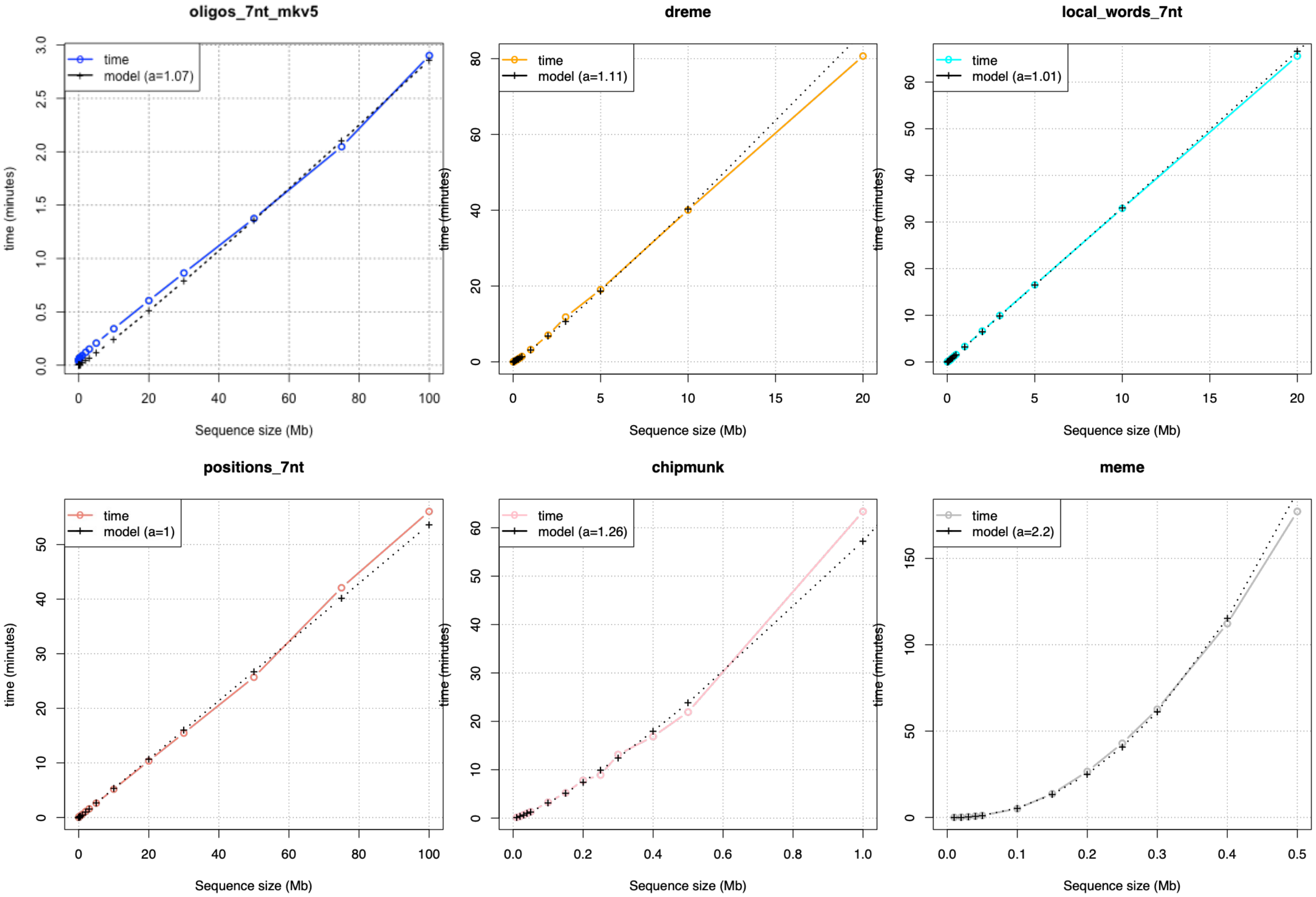
- String-based approaches
  - Time increases linearly with total sequence size.
  - Memory mostly depends on number of patterns ($N \sim 4^k$) → 100Mb can be treated with 2Gb RAM.
- EM or Gibbs samplers
  - Time >quadratic.
- On a medium-priced laptop (MacBook, 2Gb RAM), the biggest files (100Mb) is treated in
  - 3 minutes with oligo-analysis;
  - 25 minutes with dyad-analysis;
  - <1 hour with position-analysis.

Time efficiency does matter

- String-based approaches
  - **Linear complexity** : computing time proportional to total sequence size.
  - **Low memory usage** (depends on number of patterns: 4^k) → 100Mb treated with 2Gb RAM.
- EM or Gibbs samplers
  - Time >quadratic.
- On a medium-priced laptop (2010 MacBook, 2Gb RAM), the biggest files (100Mb) is treated in
  - 3 minutes with oligo-analysis;
  - 25 minutes with dyad-analysis;
  - <1 hour with position-analysis.
  - 44 years with MEME (extrapolation)

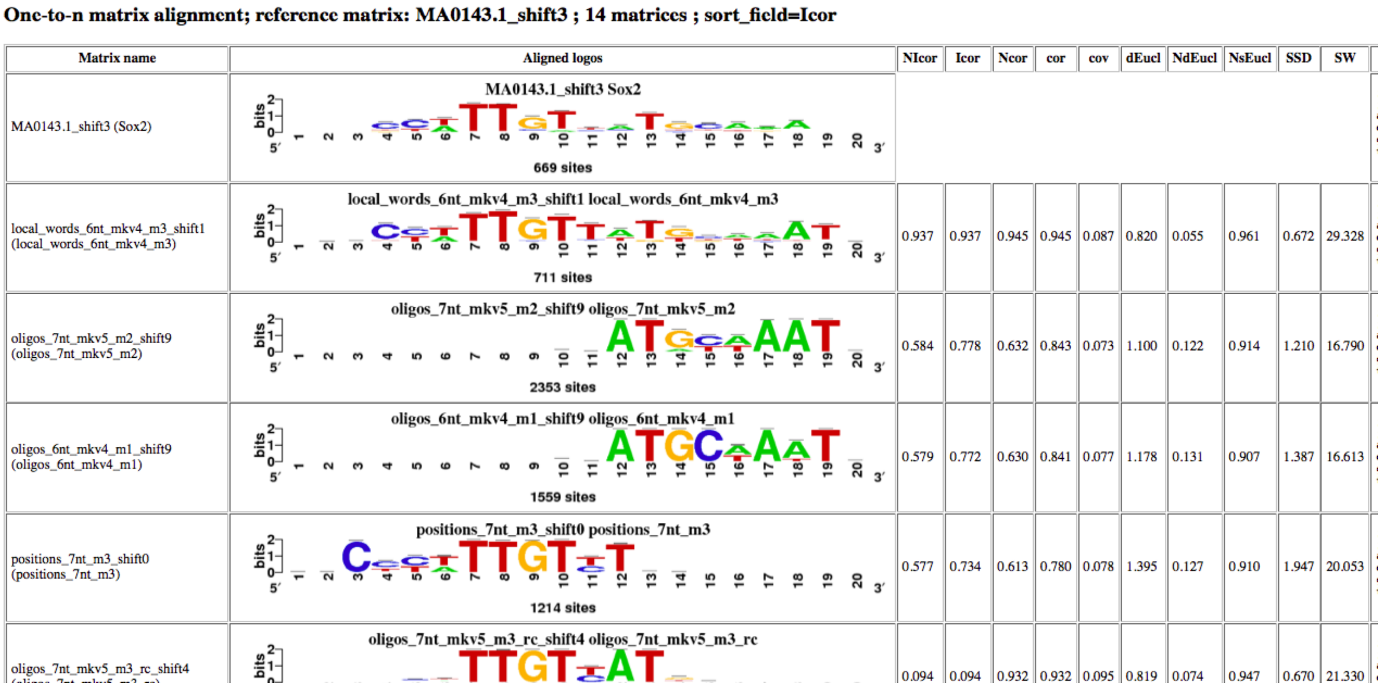


Time complexity



Discovered versus reference motifs

- Discovered motifs are compared to and aligned with the reference motifs.
- The program *compare-motifs*
  - supports various scoring schemes for assessing the similarity between motifs: correlation, Euclidian, Sandelin-Wasserman, SSD, ...
  - Generates multiple (one-to-many) alignment between matrices and logos.

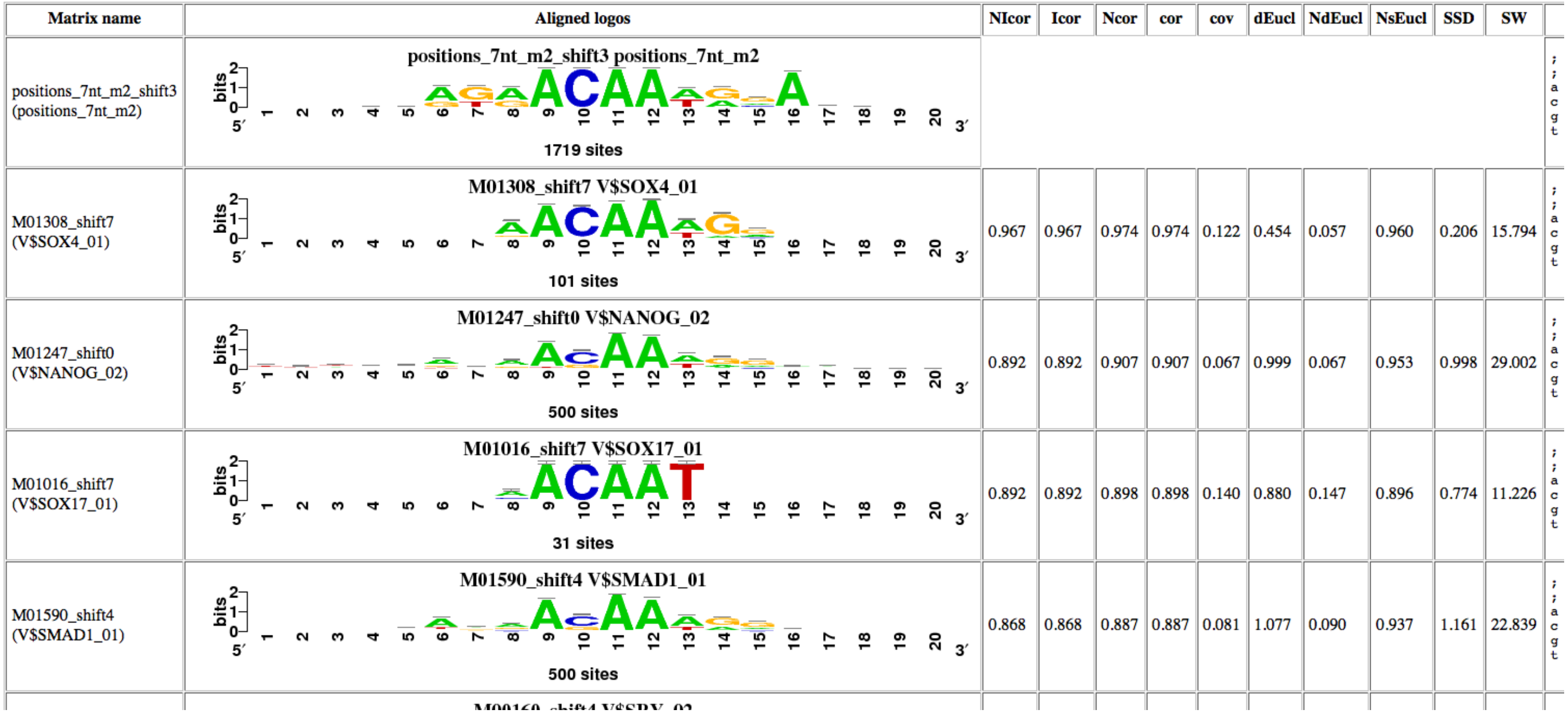


One-to-n matrix alignment; reference matrix: MA0143.1_shift3 ; 14 matrices ; sort_field=Icor

Matrix name	Aligned logos	NIcor	Icor	Ncor	cor	cov	dEucl	NdEucl	NsEucl	SSD	SW
MA0143.1_shift3 (Sox2)	MA0143.1_shift3 Sox2 669 sites										
local_words_6nt_mkv4_m3_shift1 (local_words_6nt_mkv4_m3)	local_words_6nt_mkv4_m3_shift1 local_words_6nt_mkv4_m3 711 sites	0.937	0.937	0.945	0.945	0.087	0.820	0.055	0.961	0.672	29.328
oligos_7nt_mkv5_m2_shift9 (oligos_7nt_mkv5_m2)	oligos_7nt_mkv5_m2_shift9 oligos_7nt_mkv5_m2 2353 sites	0.584	0.778	0.632	0.843	0.073	1.100	0.122	0.914	1.210	16.790
oligos_6nt_mkv4_m1_shift9 (oligos_6nt_mkv4_m1)	oligos_6nt_mkv4_m1_shift9 oligos_6nt_mkv4_m1 1559 sites	0.579	0.772	0.630	0.841	0.077	1.178	0.131	0.907	1.387	16.613
positions_7nt_m3_shift0 (positions_7nt_m3)	positions_7nt_m3_shift0 positions_7nt_m3 1214 sites	0.577	0.734	0.613	0.780	0.078	1.395	0.127	0.910	1.947	20.053
oligos_7nt_mkv5_m3_rc_shift4	oligos_7nt_mkv5_m3_rc_shift4 oligos_7nt_mkv5_m3_rc	0.094	0.094	0.932	0.932	0.095	0.819	0.074	0.947	0.670	21.330

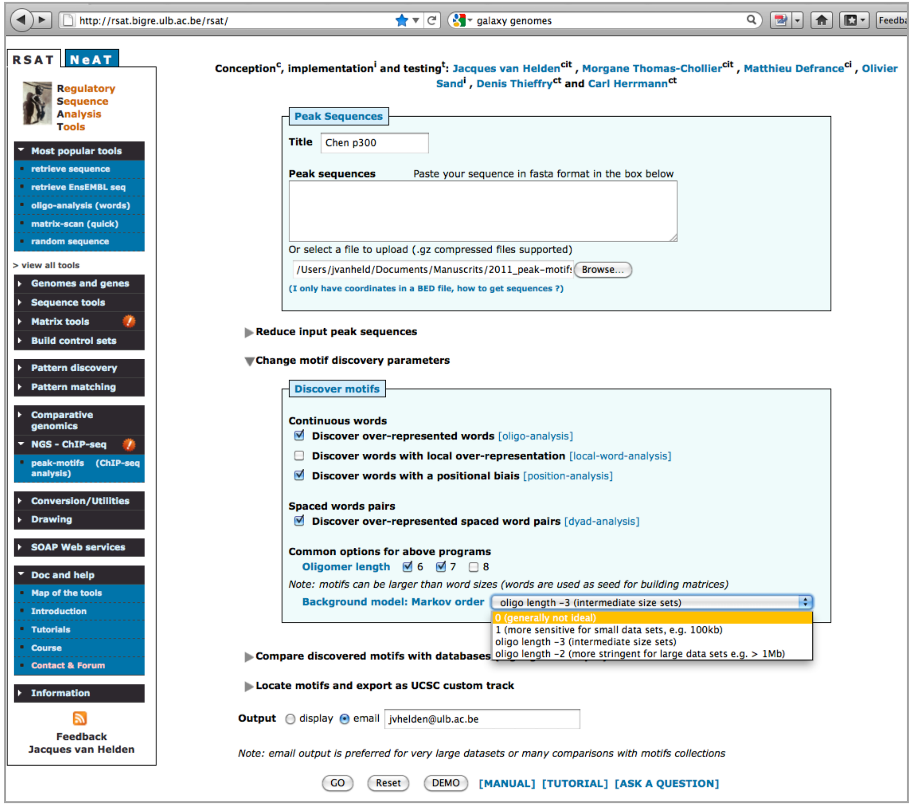
Discovered versus database motifs

- Discovered motifs are compared to all the motifs stored in specialized databases.
  - ❑ Public databases (accessible on the Web site) : JASPAR, PBM, RegulonDB, ...
  - ❑ TRANSFAC commercial database (requires local license).



Matrix name	Aligned logos	NIcor	Icor	Ncor	cor	cov	dEucl	NdEucl	NsEucl	SSD	SW	
positions_7nt_m2_shift3 (positions_7nt_m2)	positions_7nt_m2_shift3 positions_7nt_m2 1719 sites											; ; a c g t
M01308_shift7 (V$SOX4_01)	M01308_shift7 V$SOX4_01 101 sites	0.967	0.967	0.974	0.974	0.122	0.454	0.057	0.960	0.206	15.794	; ; a c g t
M01247_shift0 (V$NANOG_02)	M01247_shift0 V$NANOG_02 500 sites	0.892	0.892	0.907	0.907	0.067	0.999	0.067	0.953	0.998	29.002	; ; a c g t
M01016_shift7 (V$SOX17_01)	M01016_shift7 V$SOX17_01 31 sites	0.892	0.892	0.898	0.898	0.140	0.880	0.147	0.896	0.774	11.226	; ; a c g t
M01590_shift4 (V$SMAD1_01)	M01590_shift4 V$SMAD1_01 500 sites	0.868	0.868	0.887	0.887	0.081	1.077	0.090	0.937	1.161	22.839	; ; a c g t
	M00160_shift4 V$SRY_02											

Peak-motifs – web interface



Conception[c], implementation[i] and testing[t]: Jacques van Helden[cit], Morgane Thomas-Chollier[cit], Matthieu Defrance[ci], Olivier Sand[i], Denis Thieffry[ct] and Carl Herrmann[ct]

Peak Sequences

Title Chen p300

Peak sequences Paste your sequence in fasta format in the box below

Or select a file to upload (.gz compressed files supported)

/Users/jvanheld/Documents/Manuscrits/2011_peak-motif: Browse...

(I only have coordinates in a BED file, how to get sequences ?)

▶ Reduce input peak sequences

▼ Change motif discovery parameters

Discover motifs

Continuous words

☑ Discover over-represented words [oligo-analysis]

□ Discover words with local over-representation [local-word-analysis]

☑ Discover words with a positional biais [position-analysis]

Spaced words pairs

☑ Discover over-represented spaced word pairs [dyad-analysis]

Common options for above programs

Oligomer length ☑ 6 ☑ 7 □ 8

Note: motifs can be larger than word sizes (words are used as seed for building matrices)

Background model: Markov order: oligo length −3 (intermediate size sets)

- 0 (generally not ideal)
- 1 (more sensitive for small data sets, e.g. 100kb)
- oligo length −3 (intermediate size sets)
- oligo length −2 (more stringent for large data sets e.g. > 1Mb)

▶ Compare discovered motifs with databases

▶ Locate motifs and export as UCSC custom track

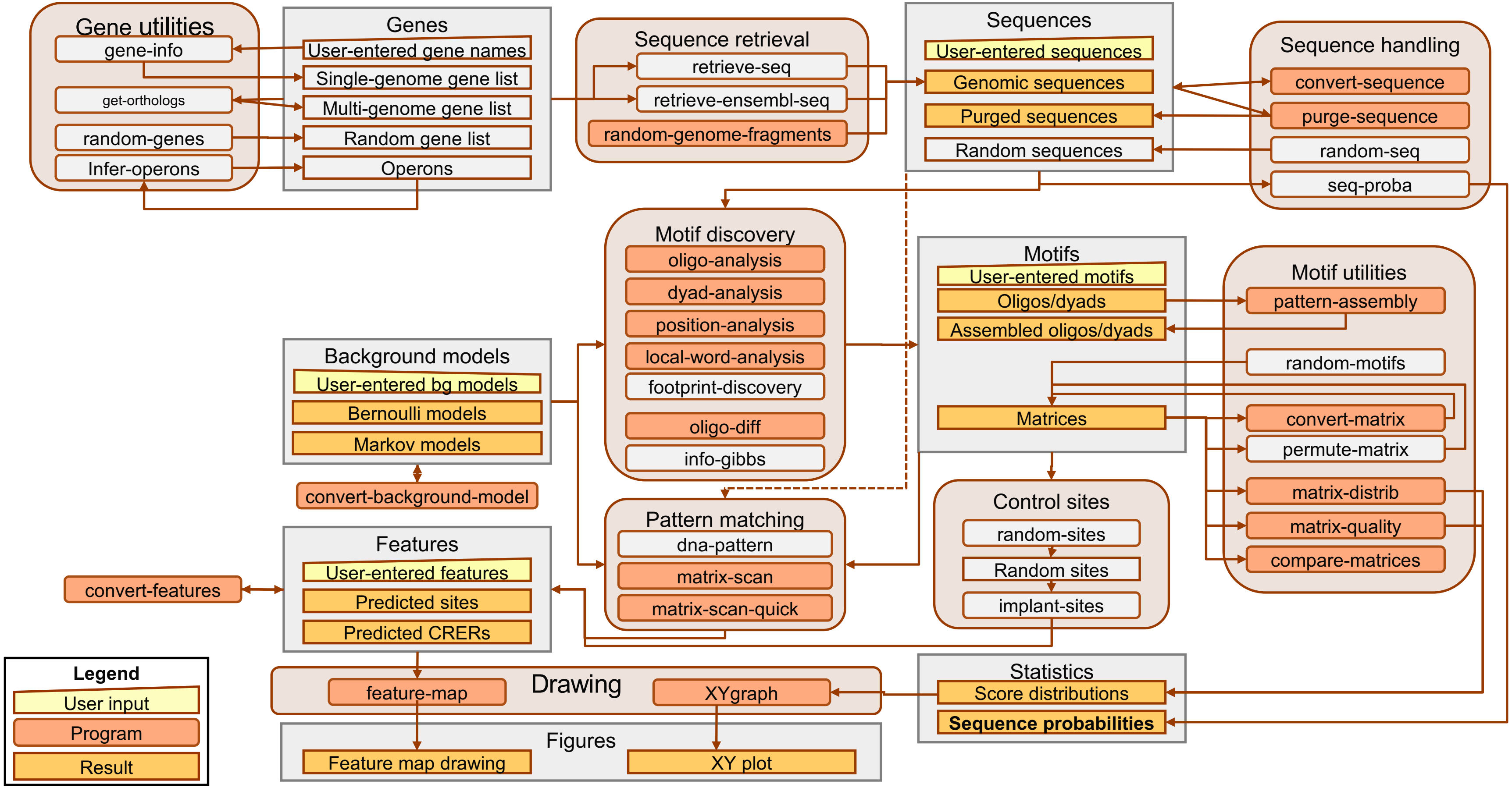
Output ○ display ● email jvhelden@ulb.ac.be

Note: email output is preferred for very large datasets or many comparisons with motifs collections

GO Reset DEMO [MANUAL] [TUTORIAL] [ASK A QUESTION]

- Regulatory Sequence Analysis Tools (RSAT)
  - http://rsat.ulb.ac.be/rsat/
- Web interface
  - Simplcity of use ("one click" interface).
  - Advanced options can be accessed optionally.
  - Allows to analyze data set of realistic size (uploaded files).
- Tutorials
- Protocol (in prep)

RSAT tools used in the peak-motifs workflow



Measuring peak enrichment in binding sites for a transcription factor of interest

Tool : matrix-quality

Building fake peak sets for negative controls

Tools:

random-seq

random-genome-fragments

random-peaks

Random-seq

- DEMO

Using motifs to evaluate peak quality

The difficulty of peak identification (peak calling)

- A ChIP-seq experiment typically returns several millions of sequences (“reads”) of short size (25bp to 100bp, depending on the sequencer characteristics).
- The reads correspond to the extremities of the DNA fragments.
  - Reads are distribued on both strands
  - The peaks on the forward and reverse strand are spaced by the average length of the fragment.
  - Most of the reads to not even cover the actual binding sites.
- Peak calling programs apply various strategies to identify and score the peaks from a set of reads, but identifying regions covered by more reads than expected by chance (see Pepke et al., 2009 for a review).
- Figure
  - RMP: read per millions.

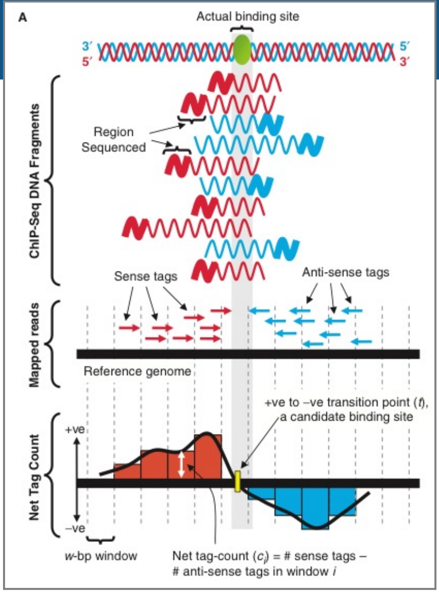
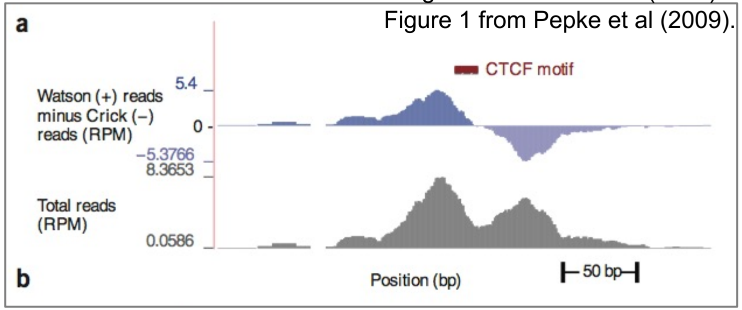


Figure from Jothi et al. (2008)

Figure 1 from Pepke et al (2009).



- Pepke et al. Computation for ChIP-seq and RNA-seq studies. Nat Methods (2009) vol. 6 (11 Suppl) pp. S22-32.
- Jothi et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res (2008) vol. 36 (16) pp. 5221-31

Question

- Which peak-caller should be used ?
  - MACS, SWEMBL, SICER, …
  - For a comparative evaluation of peak-calling programs, see Wilbanks & Facciotti (2010).
- Should we refine regions into actual peaks ?
  - PeakSplitter
- How many peaks are relevant (at least, for motif analysis)?
  - The stringency of a peak caller software strongly depends on its tuning.
    - MACS: P-value threshold (option –p, from 0 to 1)
    - SWEMBL: relative background, also called "gradient" (option –R, from 0 to 1)

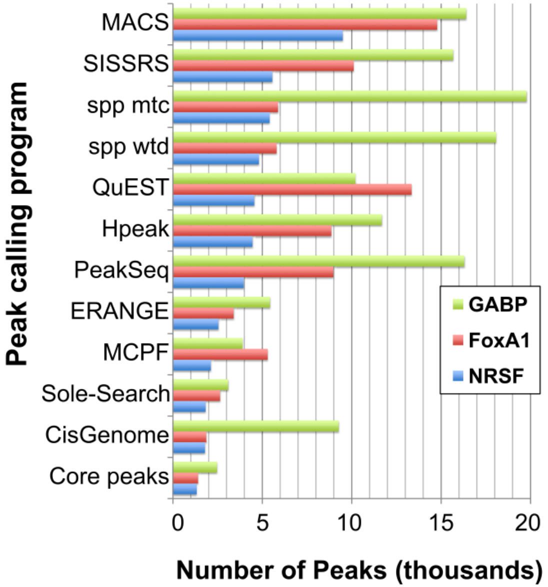
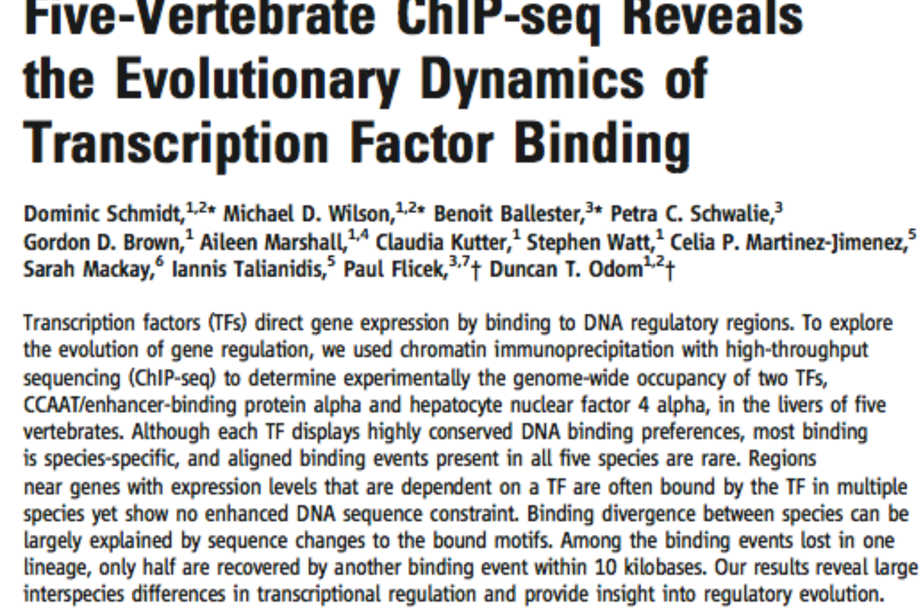


Figure 3. Quantity of peaks identified. Programs report different numbers of peaks, when run with their default or recommended settings on the same dataset. Number of reported peaks is shown for the GABP (green bars), FoxA1 (red bars) and NRSF (blue bars) datasets. To assess how different these peak lists were, those peaks identified by all 11 methods were calculated (core peaks).
doi:10.1371/journal.pone.0011471.g003

Source: Wilbanks and Facciotti (2010).

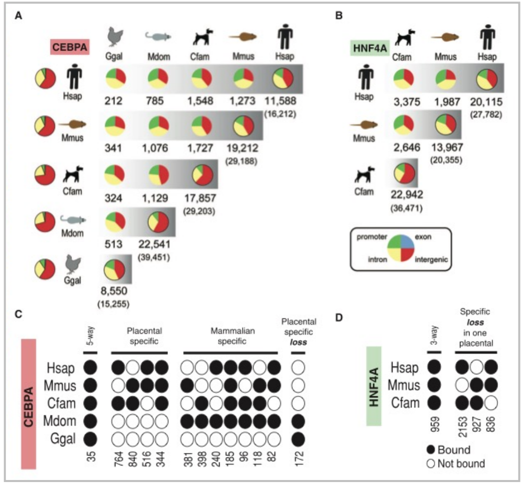
- Wilbanks EG, Facciotti MT. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. Plos One 5(7): e11471.
- Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl): S22-32.



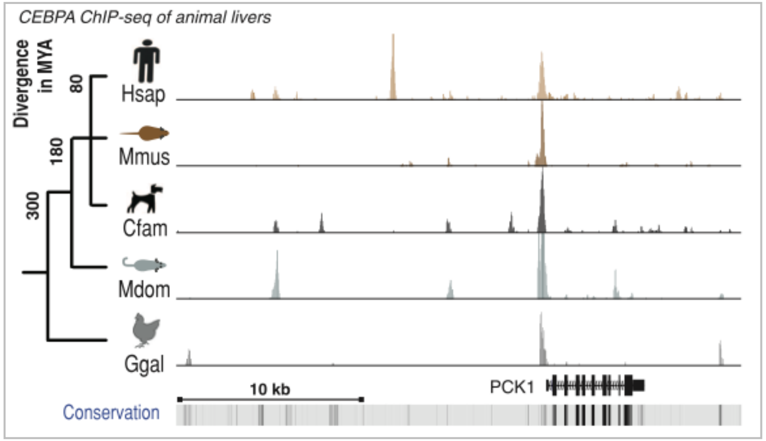
Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding

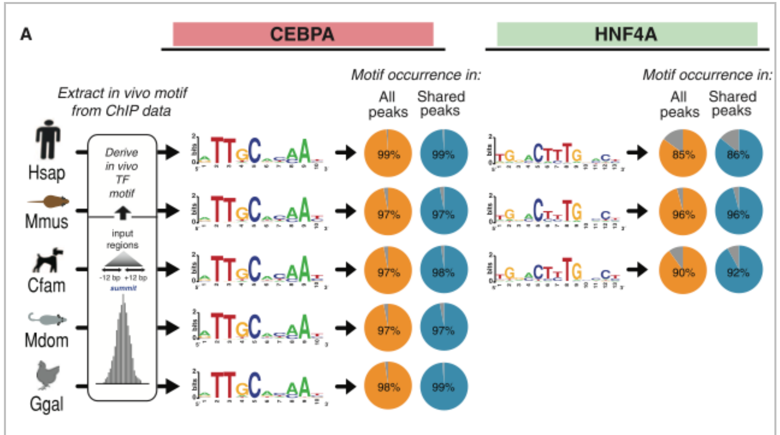
Dominic Schmidt,[1,2*] Michael D. Wilson,[1,2*] Benoit Ballester,[3*] Petra C. Schwalie,[3] Gordon D. Brown,[1] Aileen Marshall,[1,4] Claudia Kutter,[1] Stephen Watt,[1] Celia P. Martinez-Jimenez,[5] Sarah Mackay,[6] Iannis Talianidis,[5] Paul Flicek,[3,7]† Duncan T. Odom[1,2]†

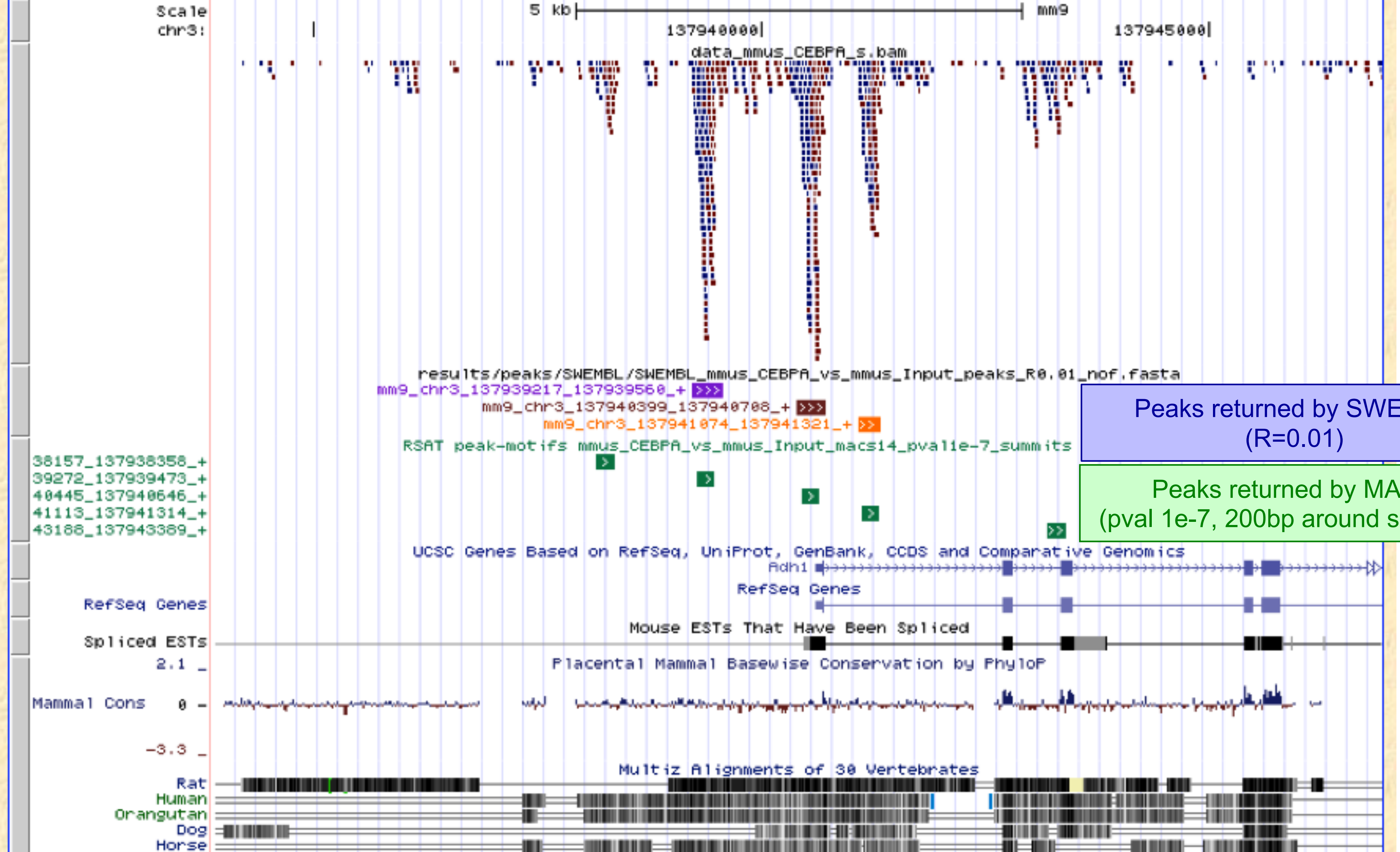
Transcription factors (TFs) direct gene expression by binding to DNA regulatory regions. To explore the evolution of gene regulation, we used chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) to determine experimentally the genome-wide occupancy of two TFs, CCAAT/enhancer-binding protein alpha and hepatocyte nuclear factor 4 alpha, in the livers of five vertebrates. Although each TF displays highly conserved DNA binding preferences, most binding is species-specific, and aligned binding events present in all five species are rare. Regions near genes with expression levels that are dependent on a TF are often bound by the TF in multiple species yet show no enhanced DNA sequence constraint. Binding divergence between species can be largely explained by sequence changes to the bound motifs. Among the binding events lost in one lineage, only half are recovered by another binding event within 10 kilobases. Our results reveal large interspecies differences in transcriptional regulation and provide insight into regulatory evolution.



Benoît Ballester (Ex-EBI; Now at TAGC)





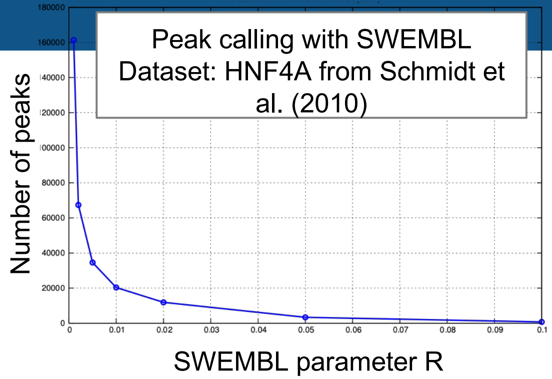

Scale
chr3:
5 kb
mm9
137940000
137945000
data_mmus_CEBPA_s.bam
results/peaks/SWEMBL/SWEMBL_mmus_CEBPA_vs_mmus_Input_peaks_R0.01_nof.fasta
mm9_chr3_137939217_137939560_+
mm9_chr3_137940399_137940708_+
mm9_chr3_137941074_137941321_+
RSAT peak-motifs mmus_CEBPA_vs_mmus_Input_macs14_pval1e-7_summits
38157_137938358_+
39272_137939473_+
40445_137940646_+
41113_137941314_+
43188_137943389_+
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Adh1
RefSeq Genes
RefSeq Genes
Mouse ESTs That Have Been Spliced
Spliced ESTs
2.1 _
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0 -
-3.3 _
Multiz Alignments of 30 Vertebrates
Rat
Human
Orangutan
Dog
Horse

Peaks returned by SWEMBL
(R=0.01)

Peaks returned by MACS
(pval 1e-7, 200bp around summits)

How many peaks?

- A non-trivial problem for the analysis of ChIP-seq data is to define the genomic regions enriched in short reads.
- The number of peaks strongly depends on the choice of the peak-calling program and on the parameters.
- Example: identifying HFN4 binding peaks in the dataset from Schmidt et al. (2010).
  - 34M reads, against 12M reads for the "input"
  - SWEMBL identifies 720 peaks when R=0.1, >160,000 peaks when R=0.001.
  - MACS identifies 52,785 peaks (parameters: mfold=10,30, pval=1e-5)
- We are getting familiar with ChIP-seq reports enumerating thousands — or tens of thousands — peaks.
  - Does it correspond to our expectation about the number of binding locations for a specific TF ?
  - How should we integrate his information in our regulatory models?



HNF4A peaks (SWEMBL)

R	# peaks	Size (Mb)
0.1	720	0.2
0.05	3,346	0.9
0.02	11,901	3.5
0.01	20,356	5.8
0.005	34,569	9.5
0.002	67,403	15.5
0.001	161,341	25.2

CEBPA peaks (SWEMBL)

R	# peaks	Size (Mb)
0.1	1,271	0.4
0.05	5,942	1.7
0.02	16,999	5.3
0.01	28,668	8.5
0.005	48,442	13.4
0.002	104,052	22.1
0.001	185,885	29.2

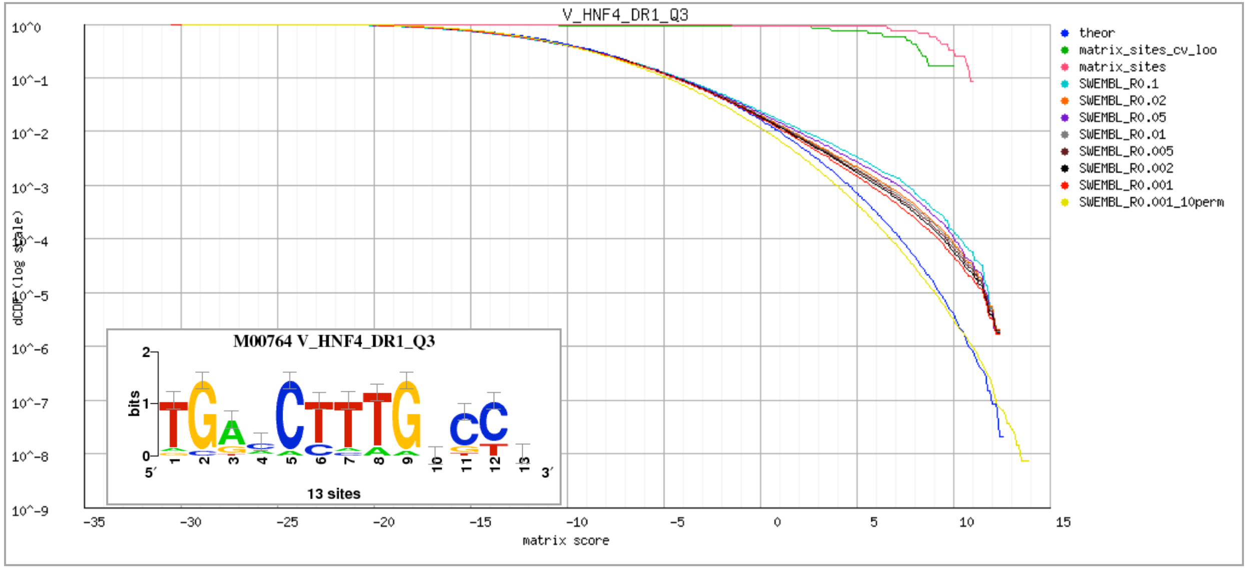
- ChIP-seq reads provided by Benoît Ballester
- Data source: Schmidt et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (2010) vol. 328 (5981) pp. 1036-40
- Tool for peak detection: SWEMBL (http://www.ebi.ac.uk/~swilder/SWEMBL/), developed by Stephan Wilder.

Peak annotation

- Context: we saw that there are many peak callers, and that the number of peaks varies depending on some parametric choices. How to evaluate the choices ?
- Significance
  - Compare discovered motif significance in actual peak with random genomic regions of the same sizes.
- Consistency
  - Compare peak sets returned by different peak callers
- Motif analysis
  - Compare the enrichment of the different peak sets for the reference motif
  - If not reference motif: compare relative enrichment for discovered motifs
- Peak annotation
  - Check the fraction of peaks in genomic regions compatible with regulation (promoters, introns)
  - Eukaryotes: compare peak sets with histone marks associated with enhancers
  - Phylogenetic conservation of predicted binding sites
- Validation
  - If a reference collection of sites is available, evaluate the overlap
  - Note: we can only evaluate the sensitivity, not the specificity since all site databases are incomplete.

Peak enrichment for the reference motif (matrix-quality)

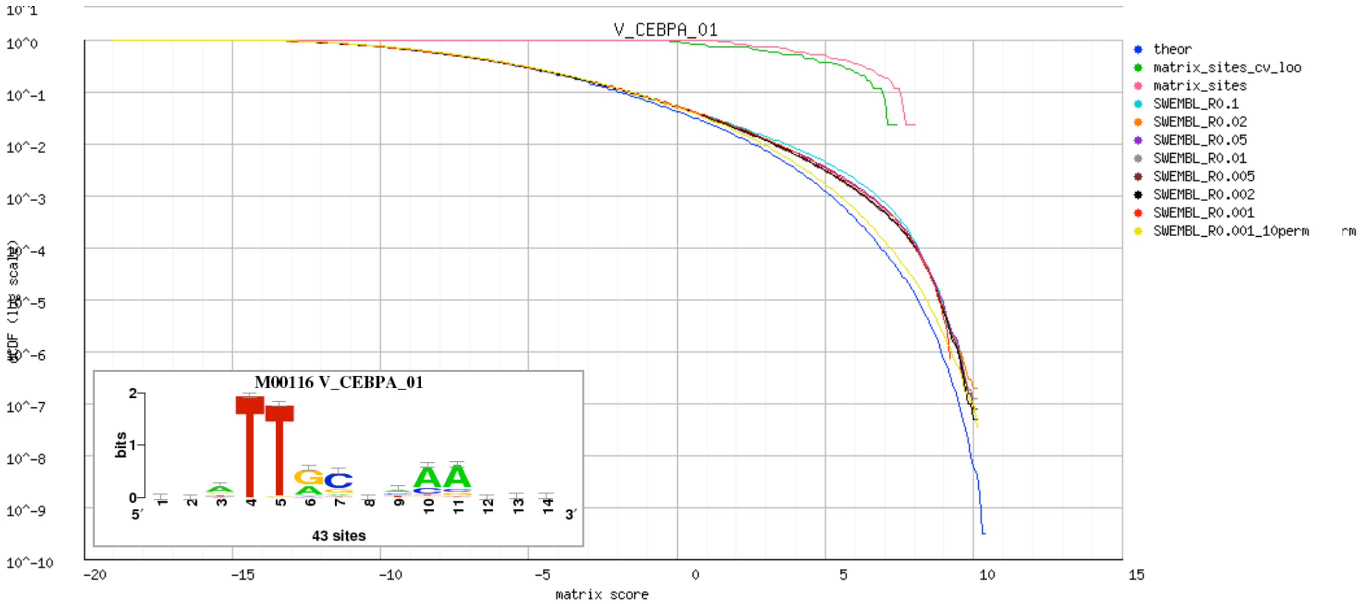
- Empirical distribution of matrix scores indicate the relative enrichment of peak sequences for the reference motif.
- Basically, the enrichment decreases when we collect more peaks (specificity decrease).
- However, we also collect more bona fide sites (sensitivity increase). 😐



- Tool : matrix-quality, Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J. & van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res 39, 808-24.
- Data source: Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328(5981): 1036-1040.

Peak enrichment for the reference motif (matrix-quality)

- Relative enrichment of peak collections for the reference motif
  - Data: With CEBPA reads from Schmidt, Wilson & Ballester
  - SWEMBL returns stronger enrichment than MACS.



- Tool : matrix-quality, Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J. & van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res 39, 808-24.
- Data source: Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328(5981): 1036-1040.

What did we learn so far ?

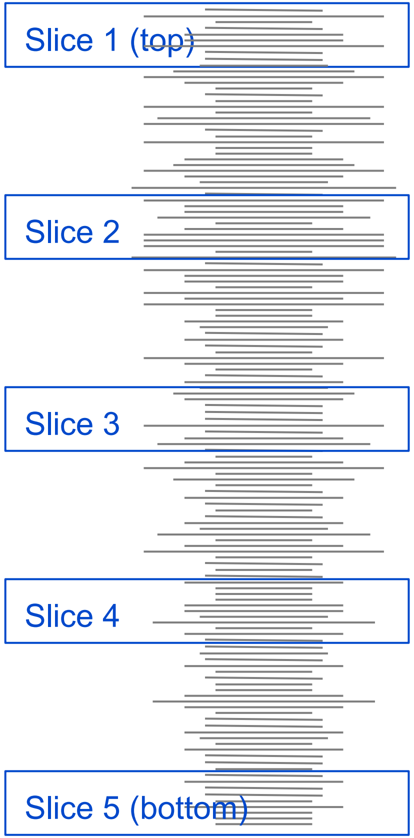
- Peak calling is not trivial
  - Peak numbers are strongly affected by the algorithm (MACS, SWEMBL, …) and parametric choices.
  - Sensitivity/specificity trade-off
    - The more peaks, the more sites.
    - But: enrichment progressively fades out.
  - All algorithms are not alike
    - With our test case, SWEMBL returns better peaks than MACS (higher enrichment for the reference motif).

- Epistemological question
  - High-throughput mind now considers normal for a TF to bind thousands, or even tens or thousands places in the genome.
  - Transcription factors may indeed spend their time to flirt with DNA, rather than bind for life to the same location.
  - However, this raises a new question: how can TFs ensure their function this way?

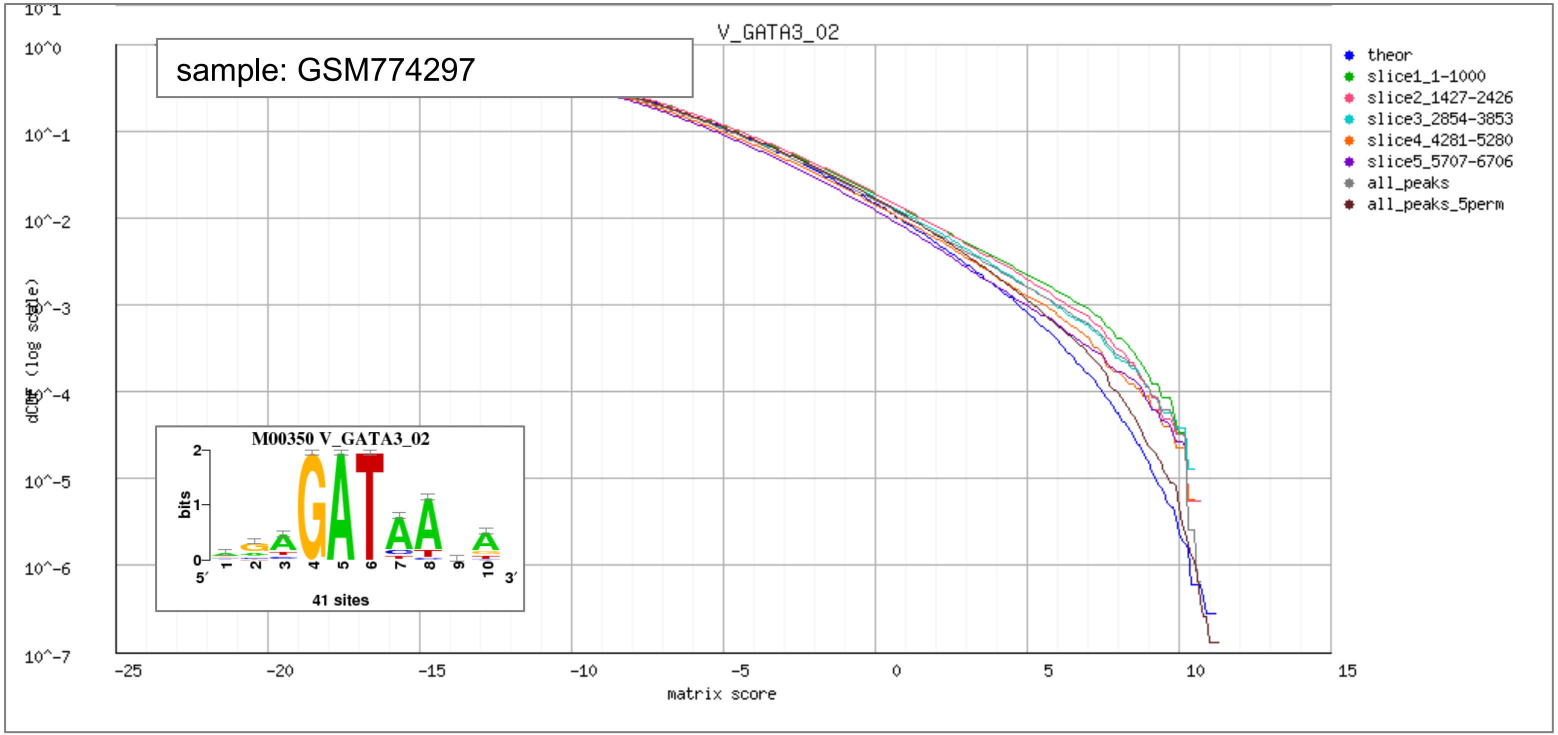
Evaluating the quality of peak collections

Slicing the peak collection

- Recipe
  - Sort peaks by decreasing score
  - Select
    - n top peaks (“top slice”)
    - n bottom peaks (“bottom slice”)
    - a few intermediate slices of n peaks
  - Analyse enrichment for a reference motif (annotated or discovered from the data) in the successive slices.

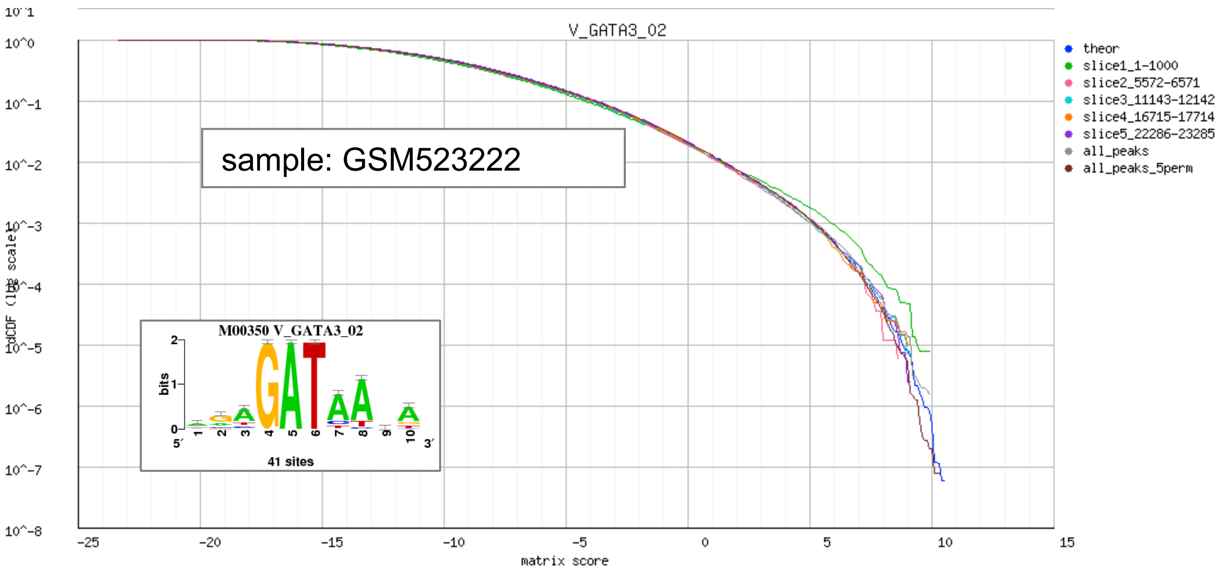


GATA3 – reasonably good peak collection



GATA3 – poor quality peak collection

- The top slice shows some enrichment
- The other slices are no more enriched than the theoretical (random) expectation
- Negative control: scanning sequences with permuted matrices fits the theoretical expectation.


V_GATA3_02
10^1
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
dCDF (log scale)
matrix score
theor
slice1_1-1000
slice2_5572-6571
slice3_11143-12142
slice4_16715-17714
slice5_22286-23285
all_peaks
all_peaks_5perm
sample: GSM523222
M00350 V_GATA3_02
bits
5′
3′
41 sites