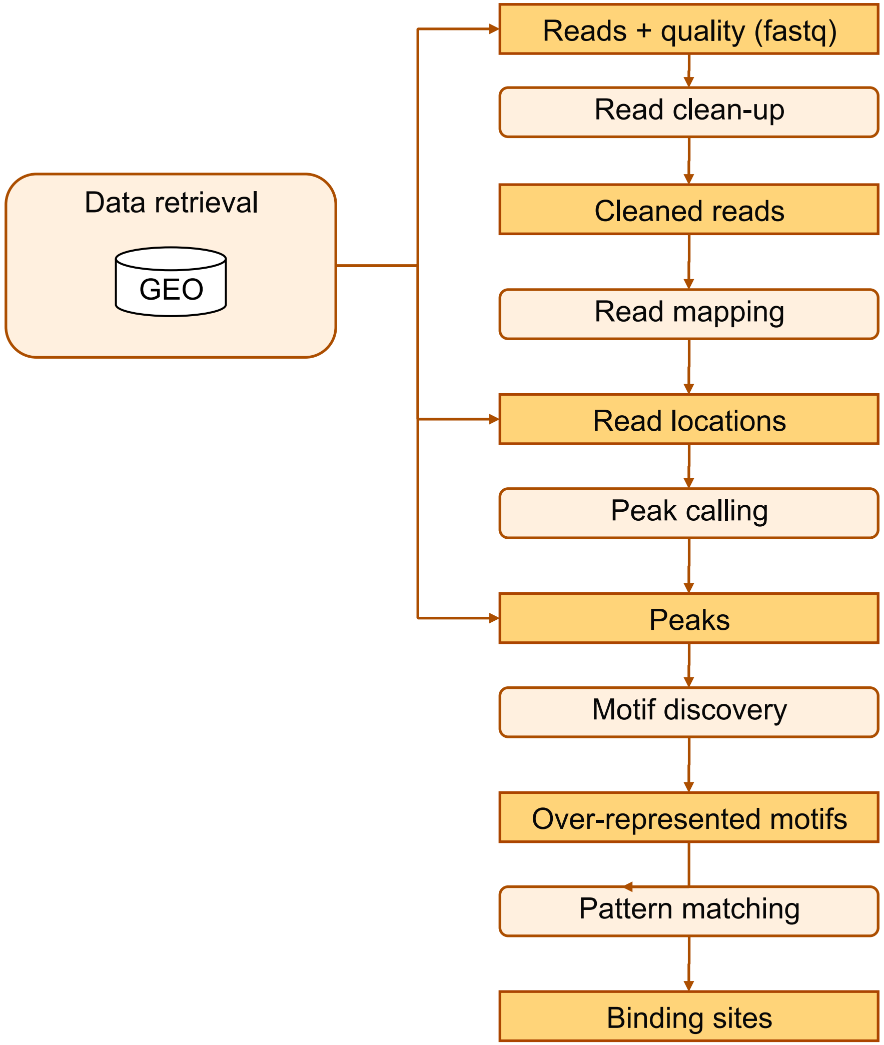
Genomics, proteomics and evolution

peak-motifs: detecting motifs in large sets of ChIP-seq peak sequences

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

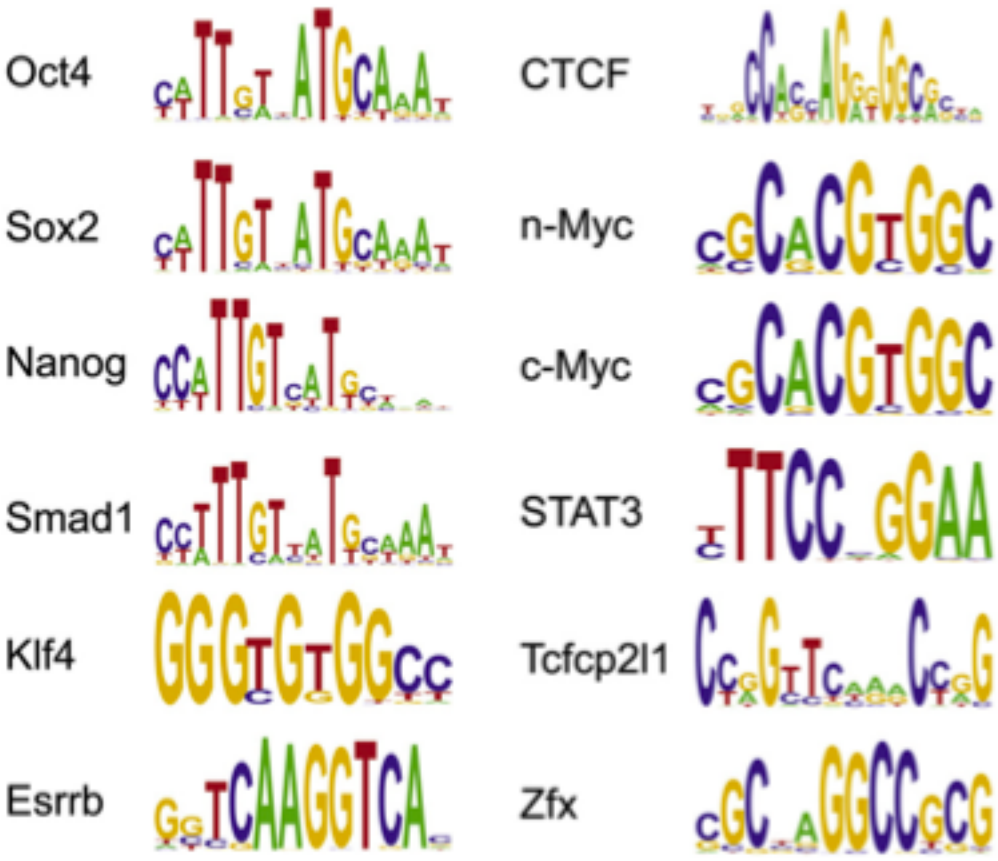
Work flow for chip-seq analysis

- ChIP-seq data can be retrieved from specialized databases such as Gene Expression Omnibus (GEO).
- The GEO database allows to retrieve sequences at various processing stages.
  - Read sequences: typically, several millions of short sequences (25bp).
  - Read locations: chromosal coordinates of each read.
  - Peak locations: several thousands of variable size regions (~10bp - 10kb).



Case study 1: Chen et al. 2008

- Binding location of 13 mouse transcription factors involved in the embryonic stem cell regulation.
- Combined the motif discovery tools Weeder and NMICA to predict motifs in each set of ChIP-seq peaks.
- Several data sets reveal the same composite motif (SOCT motif) reflecting the Sox2 / Oct4 cooperative binding.



Chen et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell (2008) vol. 133 (6) pp. 1106-17

Transcription factor binding sites: from site-wise characterization to genome-scale location (ChIP-on-chip, ChIP-seq)

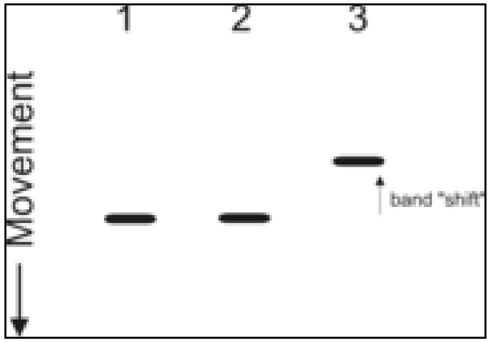
Transcription factor binding site prediction : difficulties

- Until recently, our knowledge on transcription factors relied on small collections of binding sites.
  - Such motifs are over-fitted to the few binding sites that were used to build them.

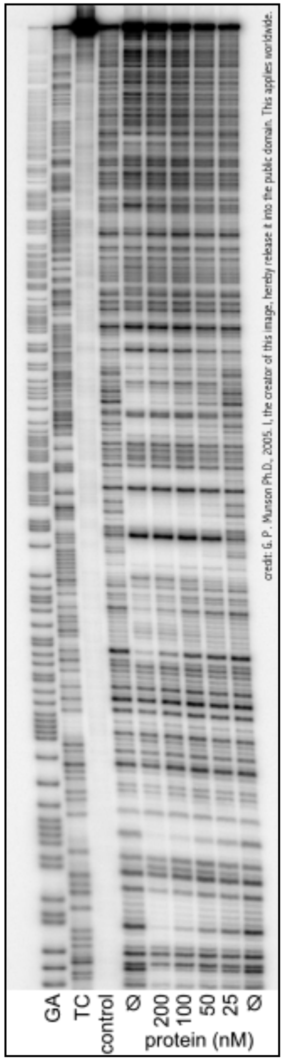
- Transcription factor binding motifs are poorly informative.
  - Motif width varies from 5 to 25 base pairs (some factors bind spaced motifs).
  - Typically 5-10 partly conserved positions.
  - Predicting individual binding sites at a genome scale is expected to return many false positives.

- The predictive power of a matrix has to be estimated on a case-by-case basis.
  - RSAT tool *matrix-quality* (Medina-Rivera et al., 2010)

Gel shift (EMSA)



DNAse footprint

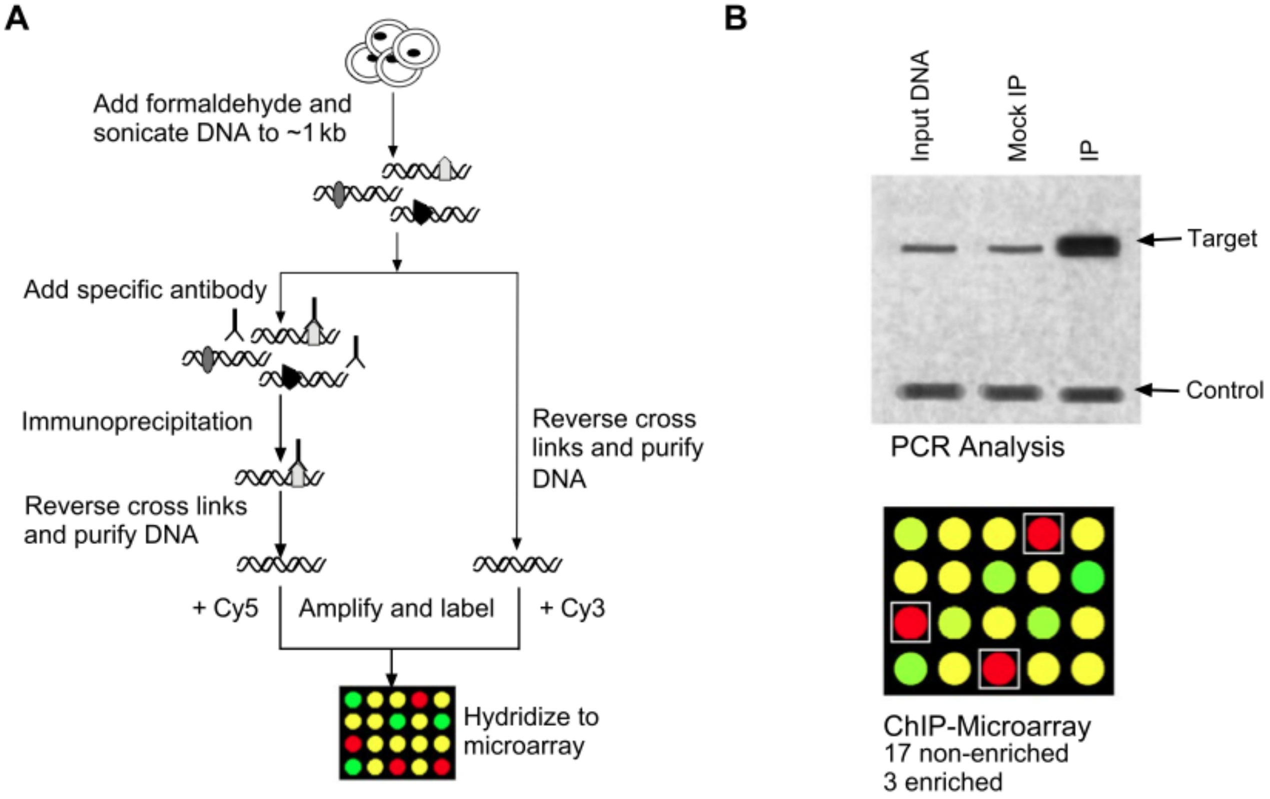


credit: G. P. Munson Ph.D., 2005. I, the creator of this image, hereby release it into the public domain. This applies worldwide.

Medina-Rivera et al. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res (2010) pp.

ChIP-on-chip

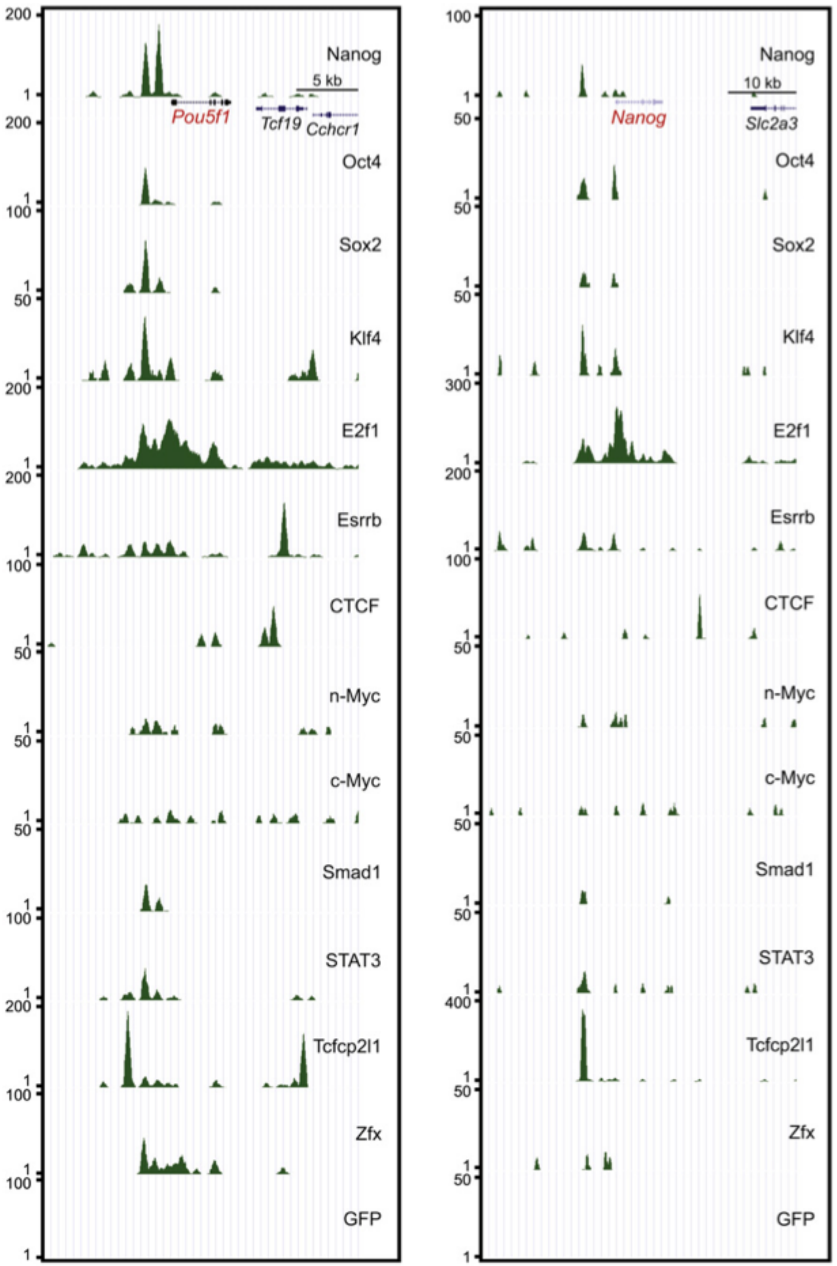
- The ChIP-on-chip method combines
  - Chromatin Immunoprecipitation (ChIP) to select genome fragments bound to a tagged transcription factor.
  - DNA microarrays (chip) spotted with several thousands of genome fragments (typically all the intergenic regions of agiven organism) are used to detect the relative enrichment: immuno-precipitated (IP) versus non-precipitated DNA (« mock » IP).
- Strength: genome-wide coverage
- Weakness: fragmentation by sonication -> large variations in DNA fragment sizes (from a few tens of bases to several kbs).



Buck and Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics (2004) vol. 83 (3) pp. 349-60

ChIP-seq

- Combination of
  - Chromatin Immunoprecipitation (ChIP), as for ChIP-chip.
  - Instead of using microarrays, the immunoprecipitated fragments are sequenced
- Strength:
  - no problem of imprecision due to the hybridation of large IP fragments to short spotted features.
  - Thanks to the « next » generation sequencing (NGC) methods, sequencing can be very efficient.
  - Does not require prior sequencing of the genome.
- Weaknesses
  - Variability of fragment sizes obtained by ultrasonication.
  - Detection of relevant peaks (peak calling) is not trivial.



Source: Chen et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell (2008) vol. 133 (6) pp. 1106-17

Read mapping

- The primary result of massively parallel sequencing is a file containing seveal millions of short sequences (the "***reads***").
- ***Read mapping*** consists in identifying the location of the reads on a genome of reference.
- This is a computationall intensive task (may take several hours on a powerful computer).

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 distributed 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.

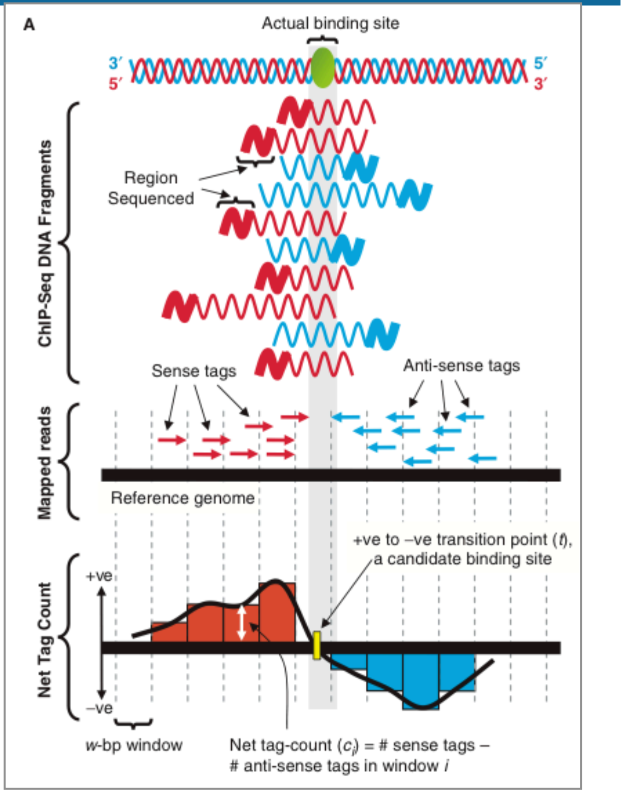


Figure from Jothi et al. (2008)

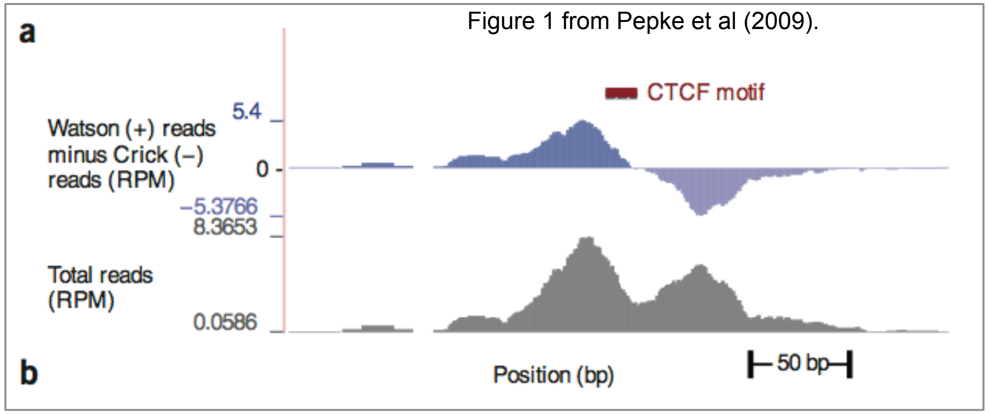
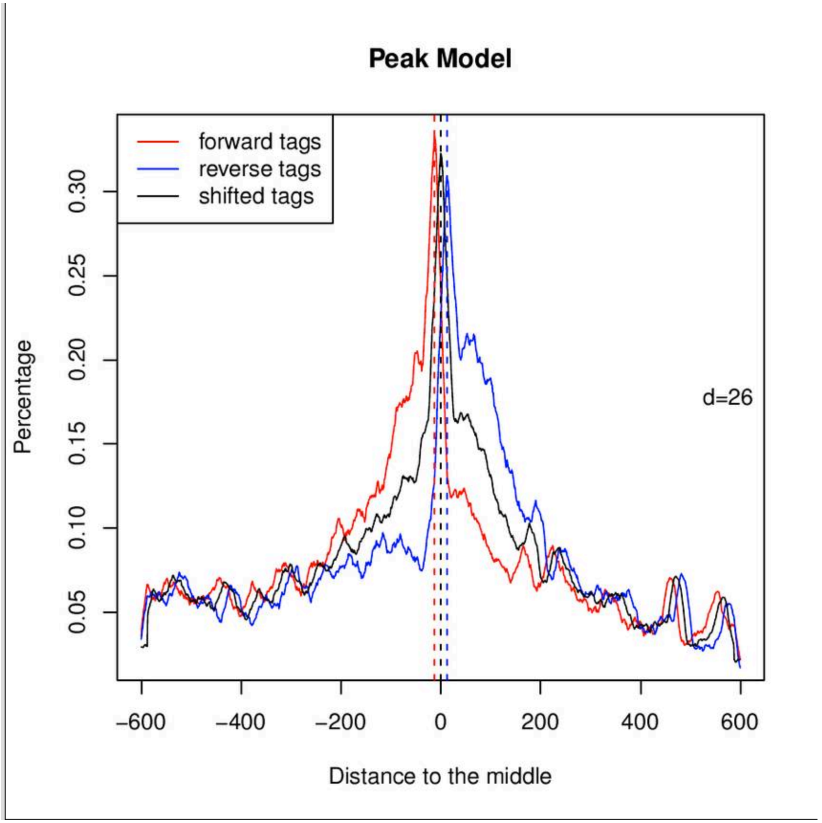


Figure 1 from Pepke et al (2009).

- 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

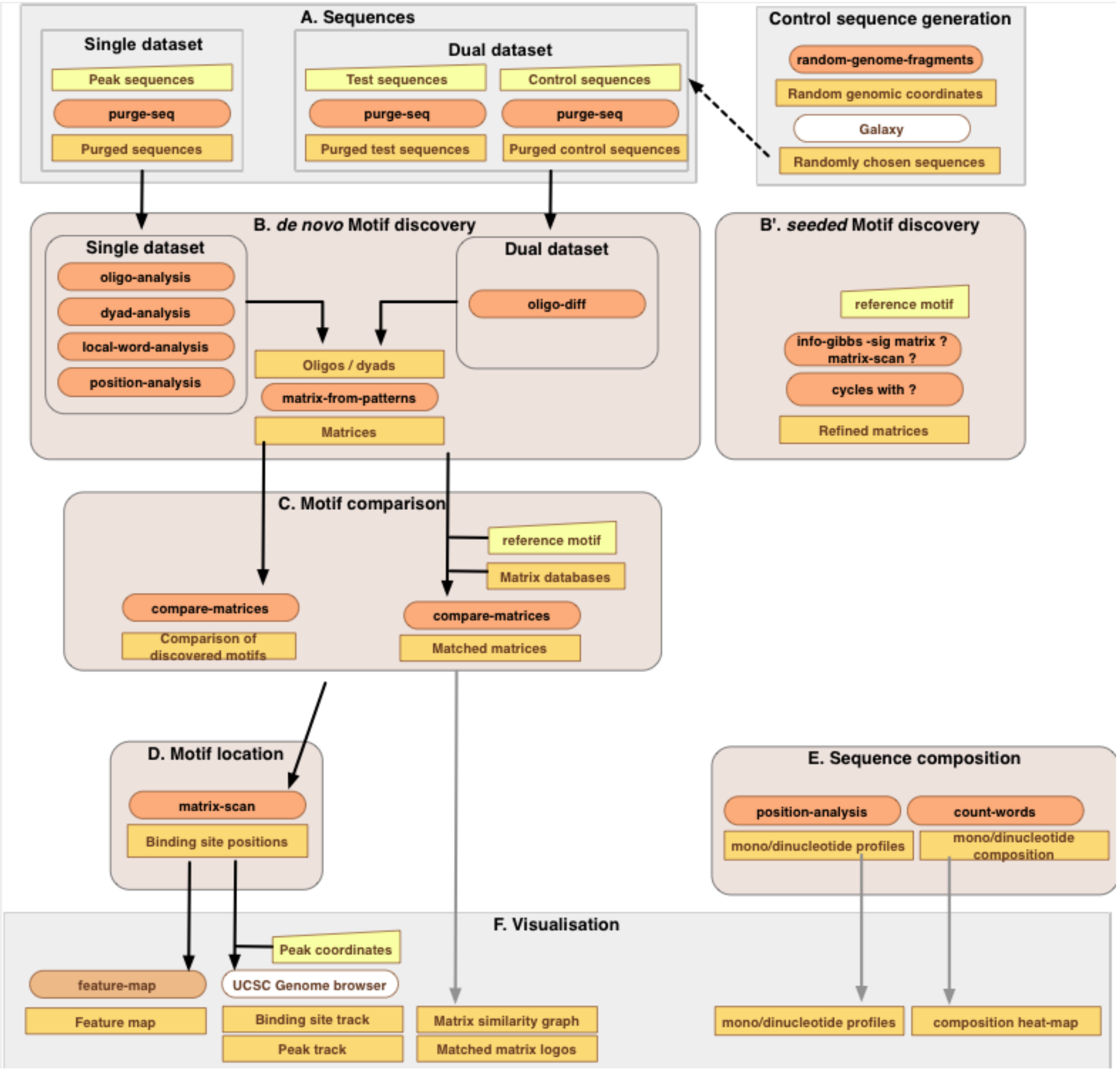
Peak calling result

- Figure: peak calling result with the reads of the Oct4 ChIP-seq from Chen 2008. Peak calling was performed with MACS on the Galaxy server.
- The curves indicate the distribution of reads relative to the centers of the peaks.
  - Red: forward strand
  - Blue: reverse strand
  - Black: "shifted" tags, obtained by comparing the forward and reverse tags.
- The 3 curves show a well-centered acute peak, which suggests that the peak calling worked well.



An integrated work flow for analyzing ChIP-seq peaks

- The program ***peak-motifs*** is a work flow that combines a series of RSAT tools in an optimal way to discoverd motifs in large sequence sets (tens of Mb) resulting from ChIP-seq experiments.
- Simple input: a set of peak sequences (fasta format).
- Multiple pattern discovery algorithms
  - Global over-representation
  - Positional biases
  - Local over-representation
- Interfaces
  - Stand-alone command
  - Web site with user-friendly interface
  - Web services (soon)



Discovering motifs in large sequence sets

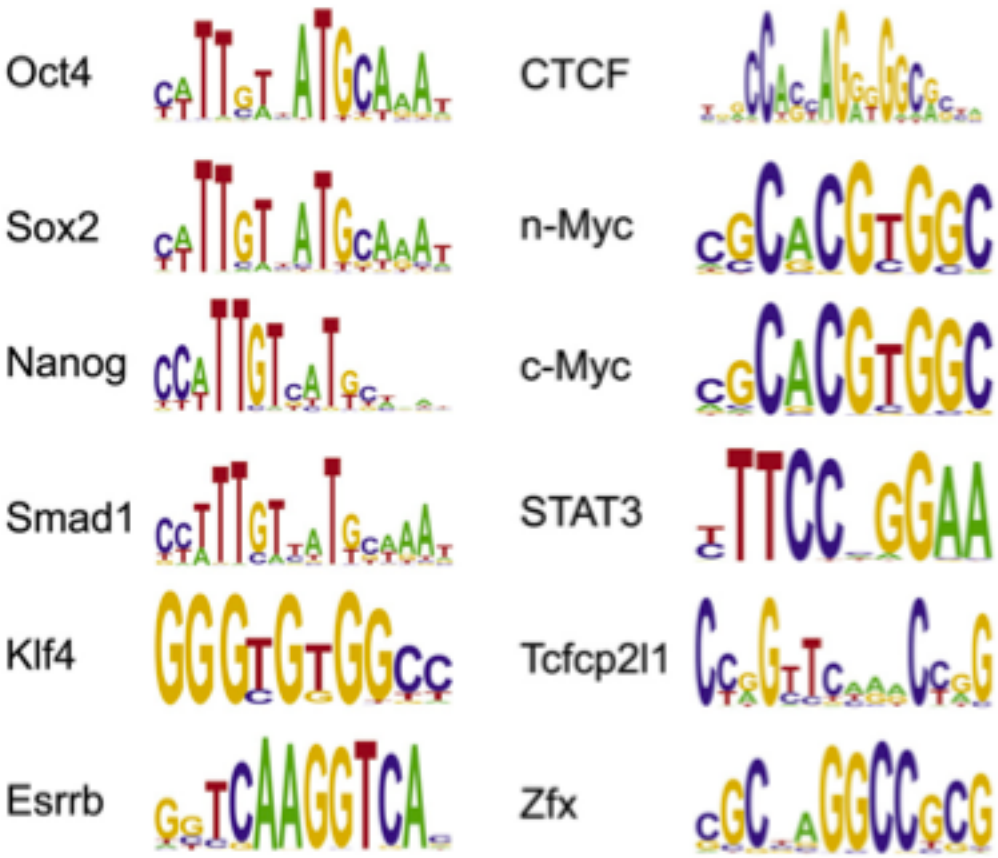
Motif discovery applied to ChIP-seq data

- Typical situation: we dispose of a collection of peak regions
  - Number : typically 1,000 to 100,000
  - Lengths: typically, between 200bp and 10,000bp, depending on
    - peak calling options
    - data type (specific transcription factor, chromatin accessibility, ...)

- Challenges
  - Extracting the “main” motif from the complete set of peak sequences (bound by the tagged TF).
  - Discovering accessory motifs (cooperative binding or frequent associations inside CRMs).
  - Comparing motifs discovered in different data sets (mutant versus WT, various conditions).
  - Predicting the precise position of binding sites inside the peak regions.

Case study 1: Chen et al. 2008

- Binding location of 13 mouse transcription factors involved in the embryonic stem cell regulation.
- Combined the motif discovery tools Weeder and NMICA to predict motifs in each set of ChIP-seq peaks.
- Several data sets reveal the same composite motif (SOCT motif) reflecting the Sox2 / Oct4 cooperative binding.



Chen et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell (2008) vol. 133 (6) pp. 1106-17

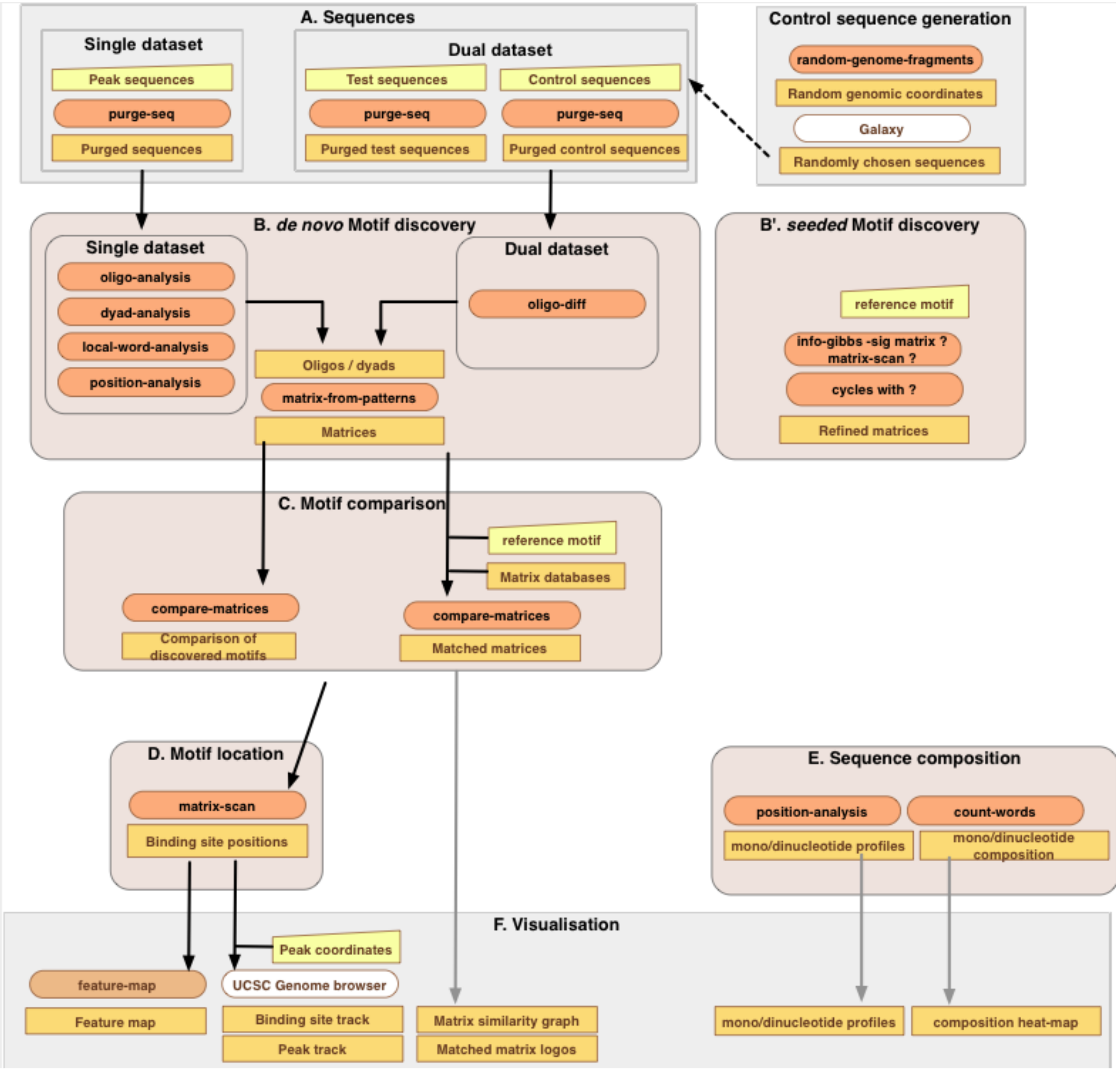
Challenges

- Motif discovery difficulties
  - Choice of the parameters for motif discovery (program, background model, ...)
- Motif discovery in peak collections is not obvious because
  - Data sets can be very large (several tens of Mb)
  - Peaks are broadly defined
  - Data sets may contain noise
  - ...

An integrated work flow for analyzing motifs in ChIP-seq and ChIP-chip peak sets

An integrated work flow for analyzing ChIP-seq peaks

- The program ***peak-motifs*** is a work flow that combines a series of RSAT tools in an optimal way to discoverd motifs in large sequence sets (tens of Mb) resulting from ChIP-seq experiments.
- Simple input: a set of peak sequences (fasta format).
- Multiple pattern discovery algorithms
  - Global over-representation
  - Positional biases
  - Local over-representation
- Interfaces
  - Stand-alone command
  - Web site with user-friendly interface
  - Web services (soon)



A. Sequences

Single dataset

Peak sequences

purge-seq

Purged sequences

Dual dataset

Test sequences

Control sequences

purge-seq

purge-seq

Purged test sequences

Purged control sequences

Control sequence generation

random-genome-fragments

Random genomic coordinates

Galaxy

Randomly chosen sequences

B. *de novo* Motif discovery

Single dataset

oligo-analysis

dyad-analysis

local-word-analysis

position-analysis

Dual dataset

oligo-diff

Oligos / dyads

matrix-from-patterns

Matrices

B'. *seeded* Motif discovery

reference motif

info-gibbs -sig matrix ?
matrix-scan ?

cycles with ?

Refined matrices

C. Motif comparison

reference motif

Matrix databases

compare-matrices

Comparison of discovered motifs

compare-matrices

Matched matrices

D. Motif location

matrix-scan

Binding site positions

E. Sequence composition

position-analysis

count-words

mono/dinucleotide profiles

mono/dinucleotide composition

F. Visualisation

Peak coordinates

feature-map

UCSC Genome browser

Feature map

Binding site track

Peak track

Matrix similarity graph

Matched matrix logos

mono/dinucleotide profiles

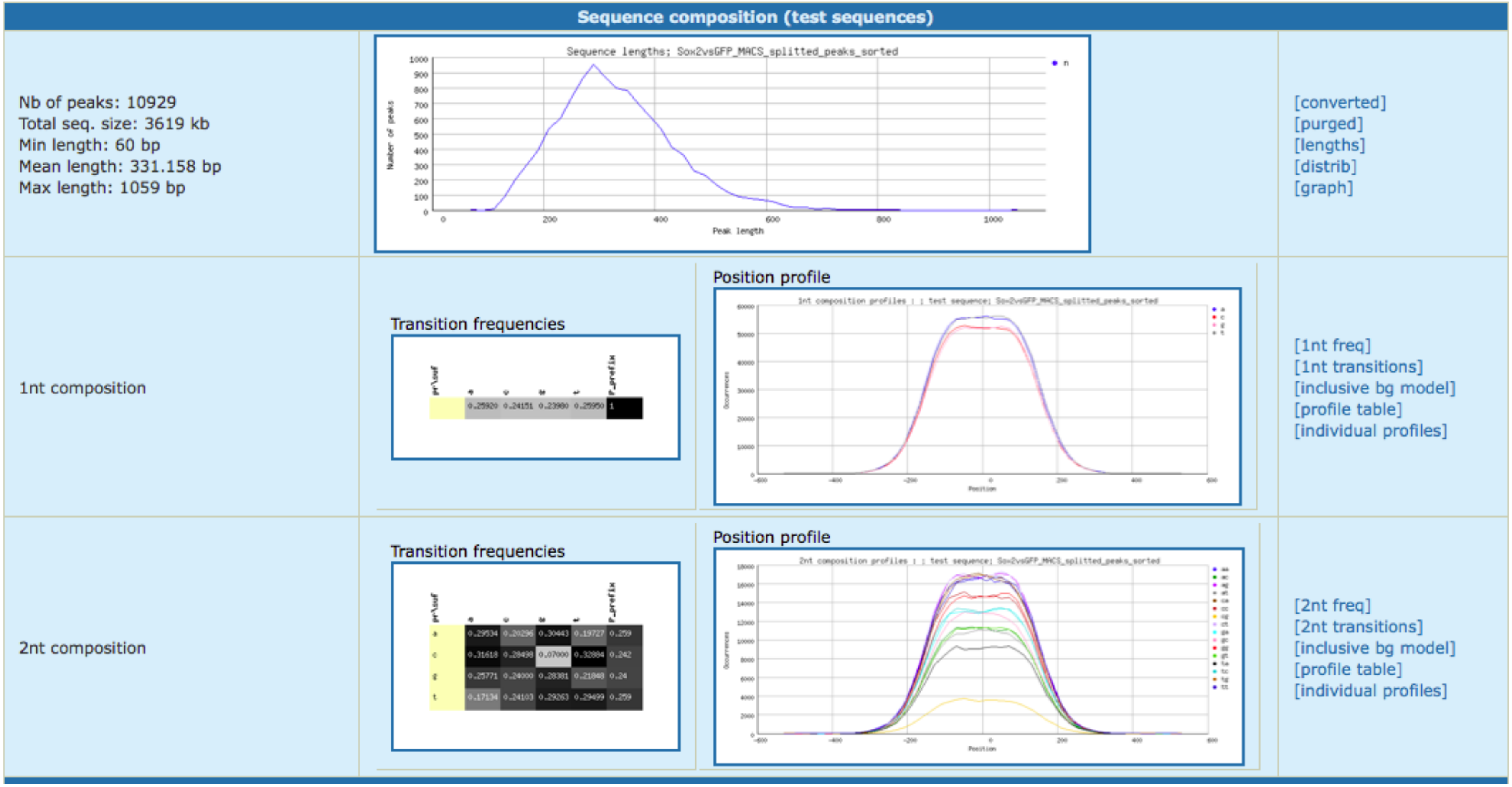
composition heat-map

Testing sets: 12 transcription factors from Chen (2008)

- Reads extracted from the GEO database
- Peaks were identified by Morgane Thomas-Chollier, using MACS on the reads, trying different options.
  - False Discovery Rate (none or 0.02)
  - Limit on the peak length
  - Peak splitting or not (split large regions into peaks)
- Reference motifs collected from TRANSFAC and JASPAR databases
  - Note: some of those motifs were obtained from high-throughput methodologies (in particular those built from the Chen dataset) -> cannot be properly speaking considered as “reference”.

Composition analysis

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



Sequence composition (test sequences)

Nb of peaks: 10929
Total seq. size: 3619 kb
Min length: 60 bp
Mean length: 331.158 bp
Max length: 1059 bp

Sequence lengths; Sox2vsGFP_MACS_splitted_peaks_sorted
Number of peaks
Peak length
n
0
100
200
300
400
500
600
700
800
900
1000

[converted]
[purged]
[lengths]
[distrib]
[graph]

1nt composition

Transition frequencies

pr\suf	a	c	g	t	P_prefix
	0.25920	0.24151	0.23980	0.25950	1

Position profile

1nt composition profiles : : test sequence: Sox2vsGFP_MACS_splitted_peaks_sorted
Occurrences
Position
a
c
g
t
-600
-400
-200
0
200
400
600

[1nt freq]
[1nt transitions]
[inclusive bg model]
[profile table]
[individual profiles]

2nt composition

Transition frequencies

pr\suf	a	c	g	t	P_prefix
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

Position profile

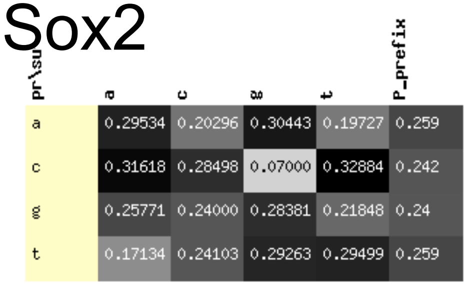
2nt composition profiles : : test sequence: Sox2vsGFP_MACS_splitted_peaks_sorted
Occurrences
Position
-600
-400
-200
0
200
400
600

[2nt freq]
[2nt transitions]
[inclusive bg model]
[profile table]
[individual profiles]

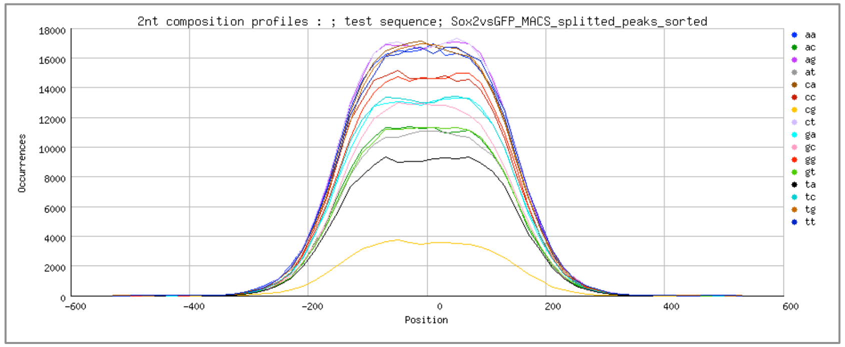
Composition analysis results

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

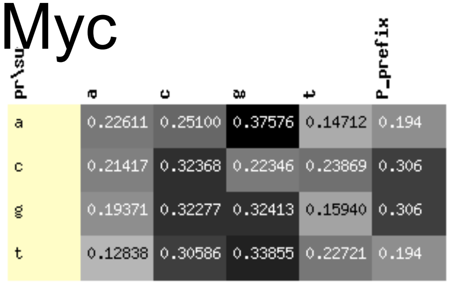
Sox2



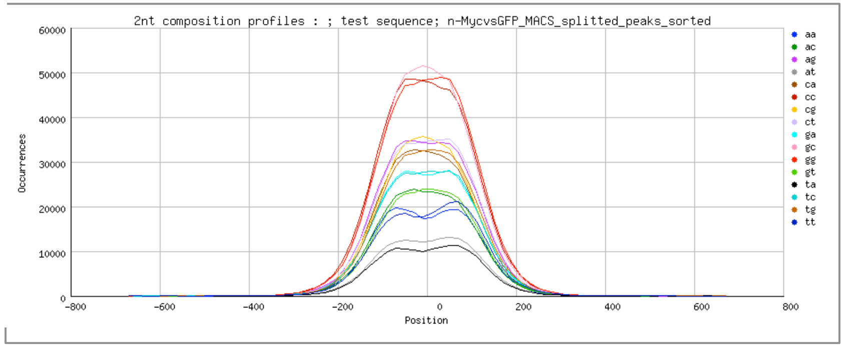
pr\su	a	c	g	t	P_prefix
a	0.29534	0.20296	0.30443	0.19727	0.259
c	0.31618	0.28498	0.07000	0.32884	0.242
g	0.25771	0.24000	0.28381	0.21848	0.24
t	0.17134	0.24103	0.29263	0.29499	0.259



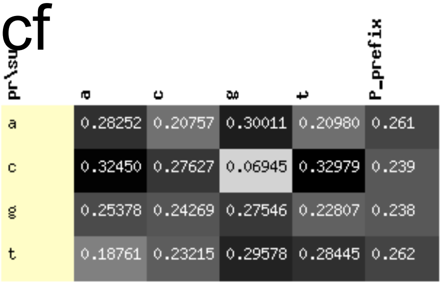
n-Myc



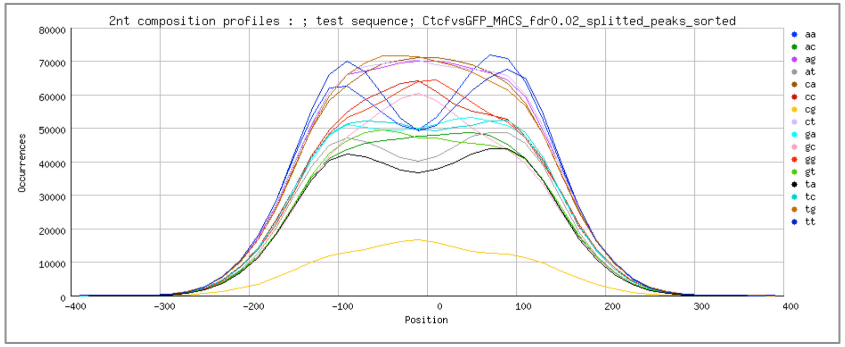
pr\su	a	c	g	t	P_prefix
a	0.22611	0.25100	0.37576	0.14712	0.194
c	0.21417	0.32368	0.22346	0.23869	0.306
g	0.19371	0.32277	0.32413	0.15940	0.306
t	0.12838	0.30586	0.33855	0.22721	0.194



Ctcf



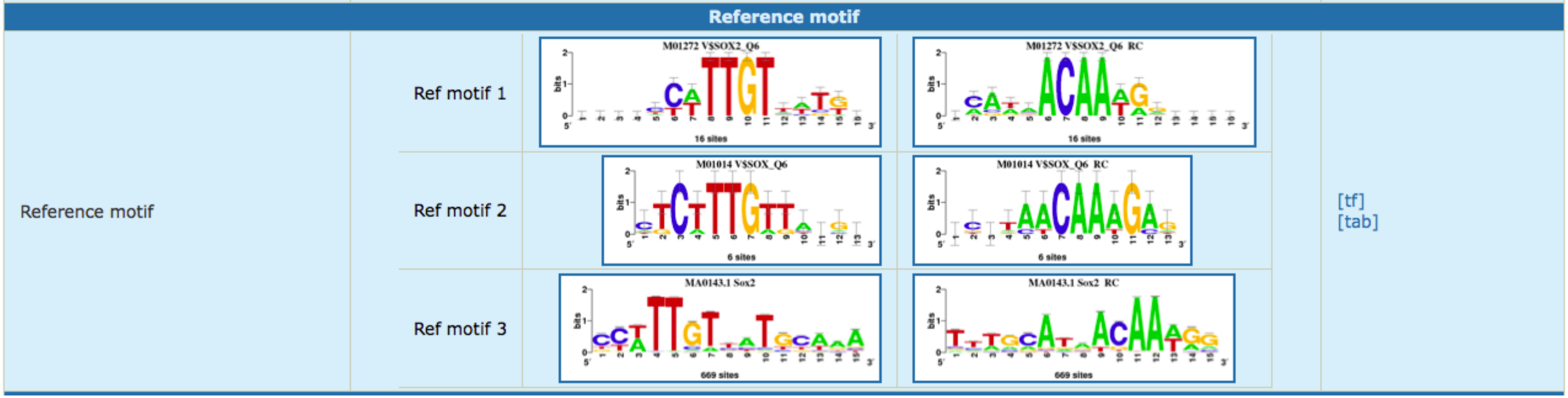
pr\su	a	c	g	t	P_prefix
a	0.28252	0.20757	0.30011	0.20980	0.261
c	0.32450	0.27627	0.06945	0.32979	0.239
g	0.25378	0.24269	0.27546	0.22807	0.238
t	0.18761	0.23215	0.29578	0.28445	0.262



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

Reference motifs

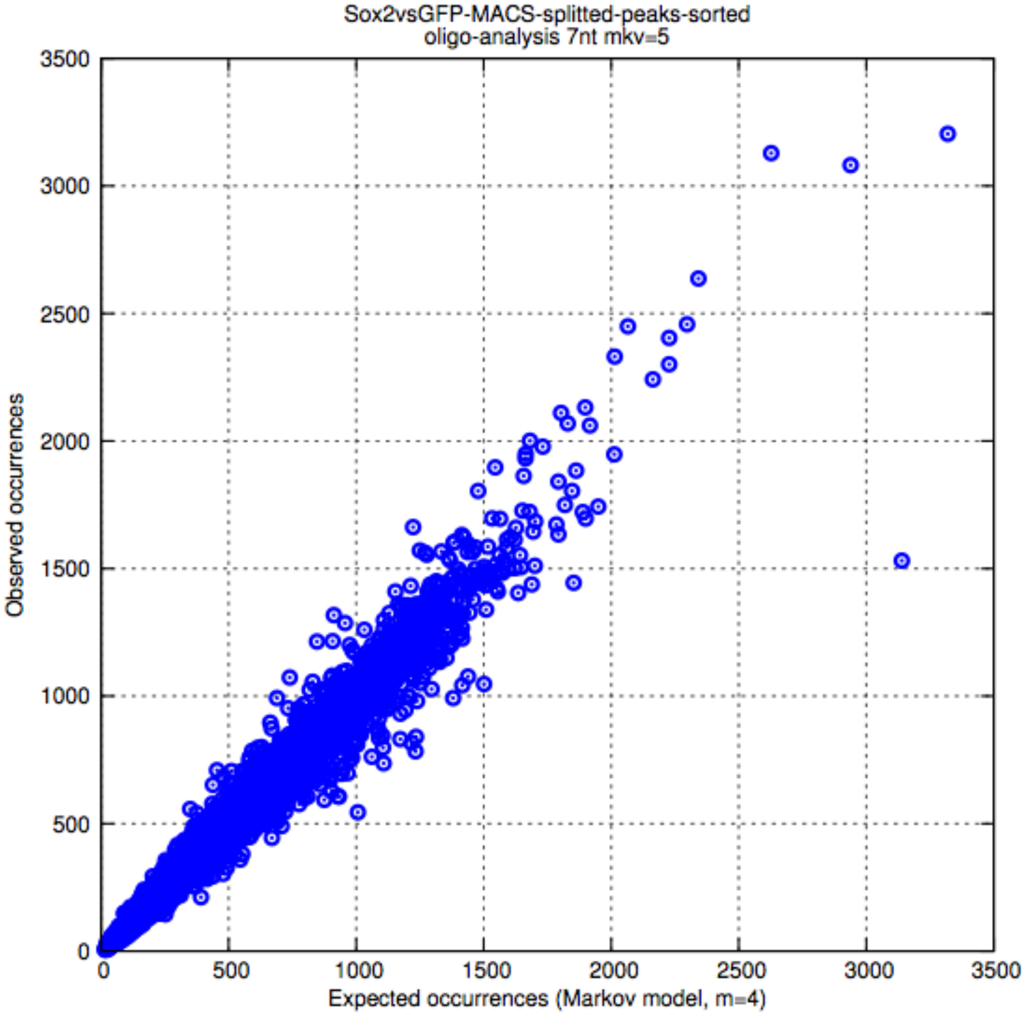
- One or several reference motifs can be defined.
- Reference motifs are the ones which are expected to be found in the dataset.
    - More precisely, if those motifs are not reported, it is considered as a failure.
- Choice of reference motifs is somewhat tricky.
    - Ex: Sox2 peaks
    - 2 slightly 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				
Reference motif	Ref motif 1	M01272 V$SOX2_Q6	M01272 V$SOX2_Q6 RC	[tf] [tab]
	Ref motif 2	M01014 V$SOX_Q6	M01014 V$SOX_Q6 RC	
	Ref motif 3	MA0143.1 Sox2	MA0143.1 Sox2 RC	

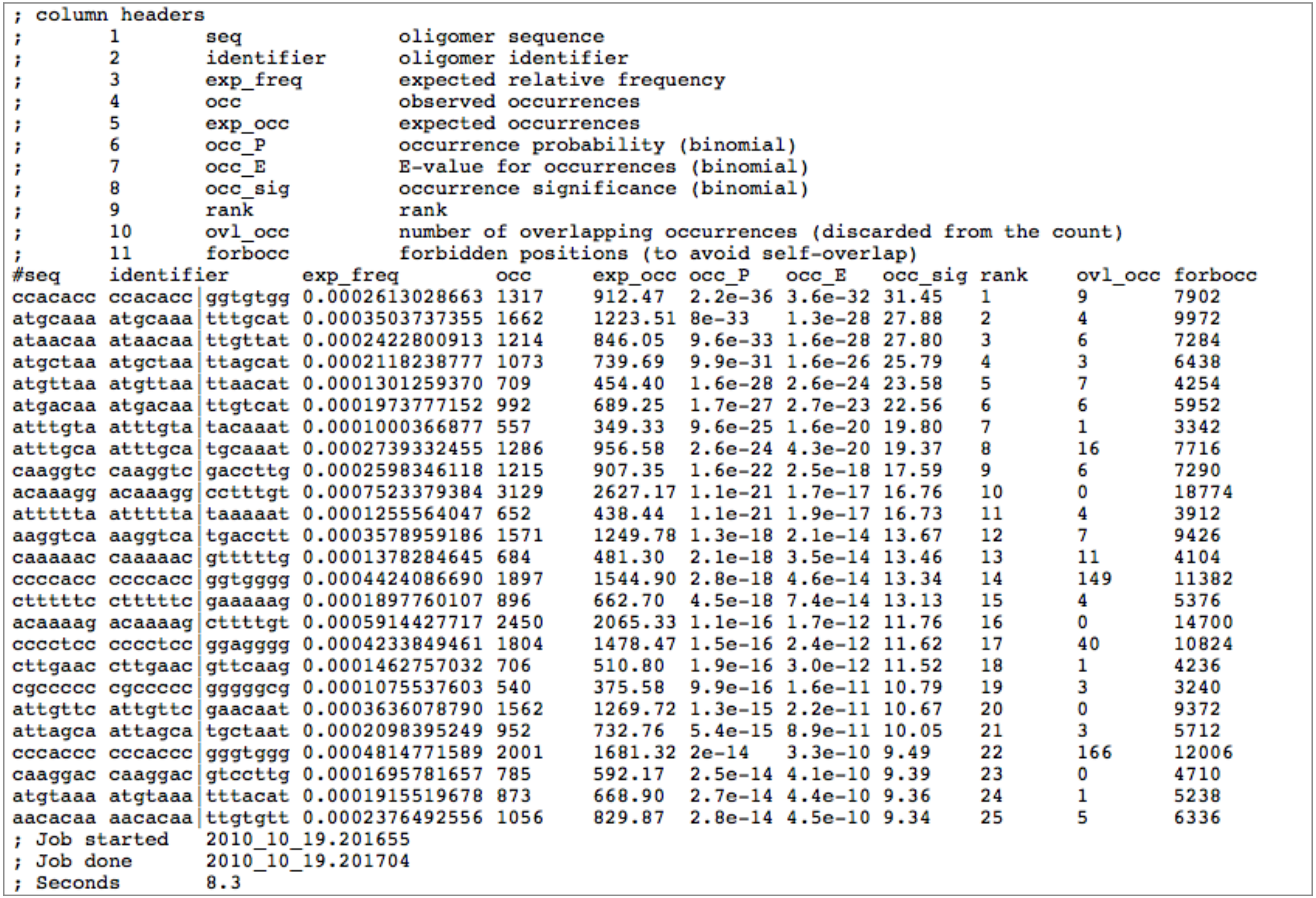
Detection of over-represented oligonucleotides (oligo-analysis)

- Principle
  - Count the occurrences of all words (oligonucleotides) of a given size in the input set
  - Estimate the expected number of occurrences according to some background model
  - Report significantly over-represented words.
- Example
  - Sox2 peaks from Chen (2008).
  - Word length *k*=7
  - Markov model of order m=5 trained on the input set.



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.

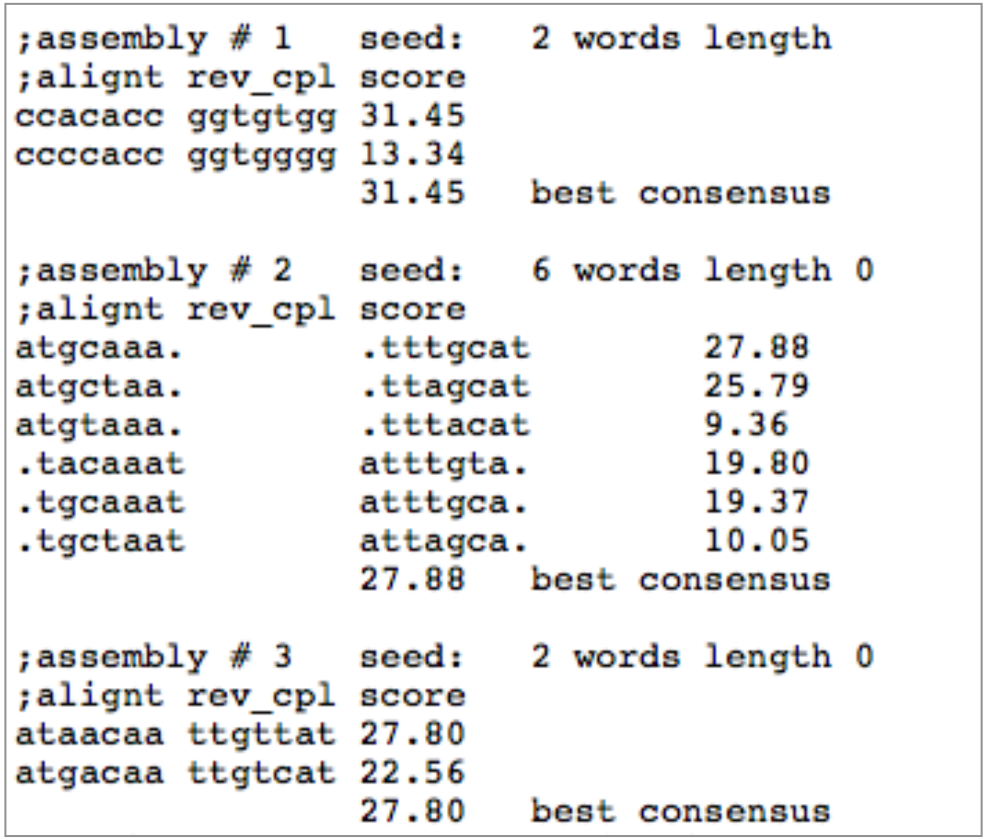
Primary result: a list of over-represented words



```
; 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
attttta attttta|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
```

The over-represented words can be assembled

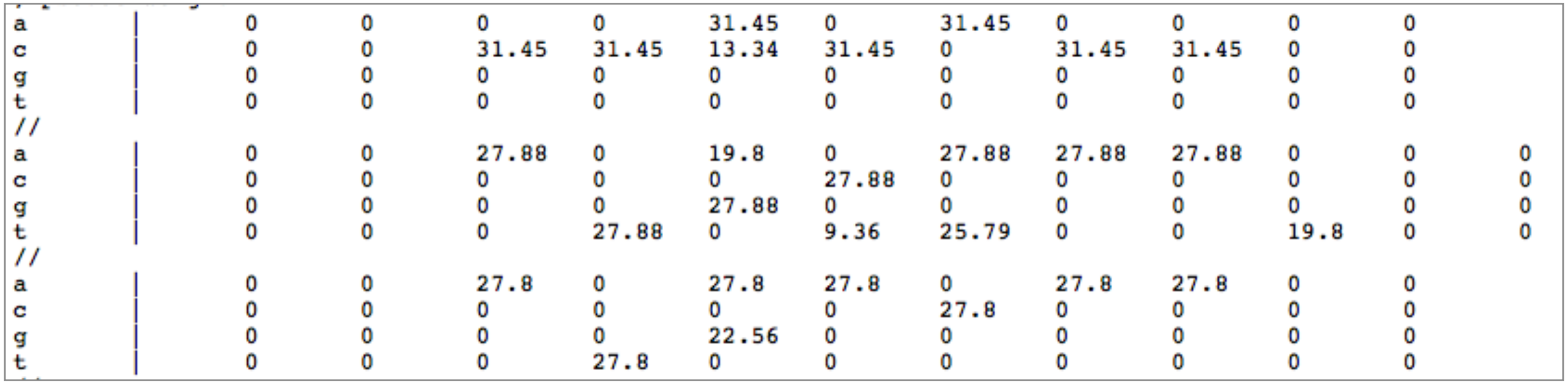
- The list of over-represented words generally contain groups of mutually overlapping words.
- Those groups can be aligned using the program *pattern-assembly*
- Assembled words reveal
  - larger motifs than the initial word length
  - positions with variable residues
- Word assemblies can be used to build a significance matrix (example below).



```
;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
```

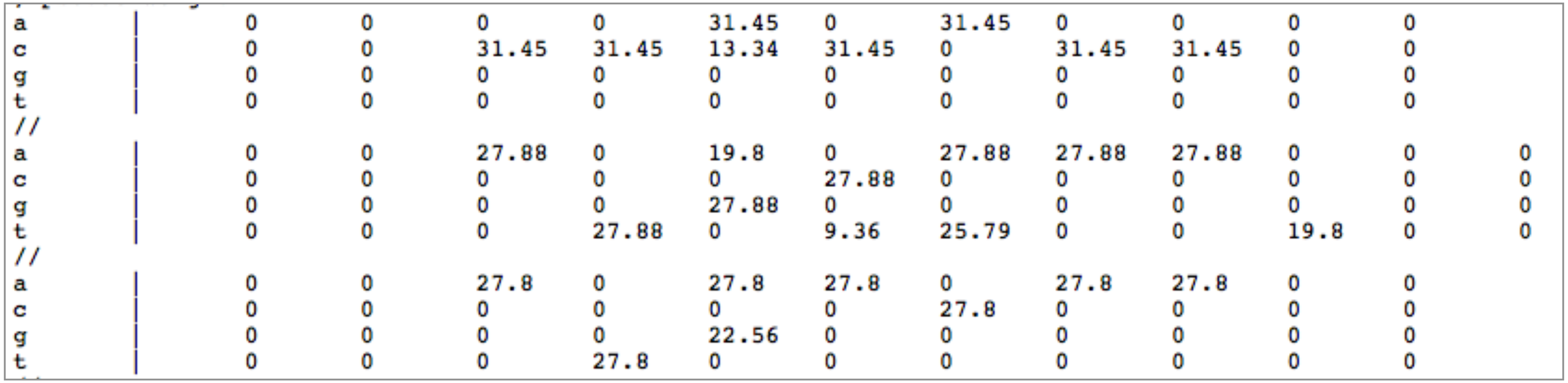


```
a       |       0       0       0       0       31.45   0       31.45   0       0       0       0
c       |       0       0       31.45   31.45   13.34   31.45   0       31.45   31.45   0       0
g       |       0       0       0       0       0       0       0       0       0       0       0
t       |       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
```

Collecting a matrix from assembled words

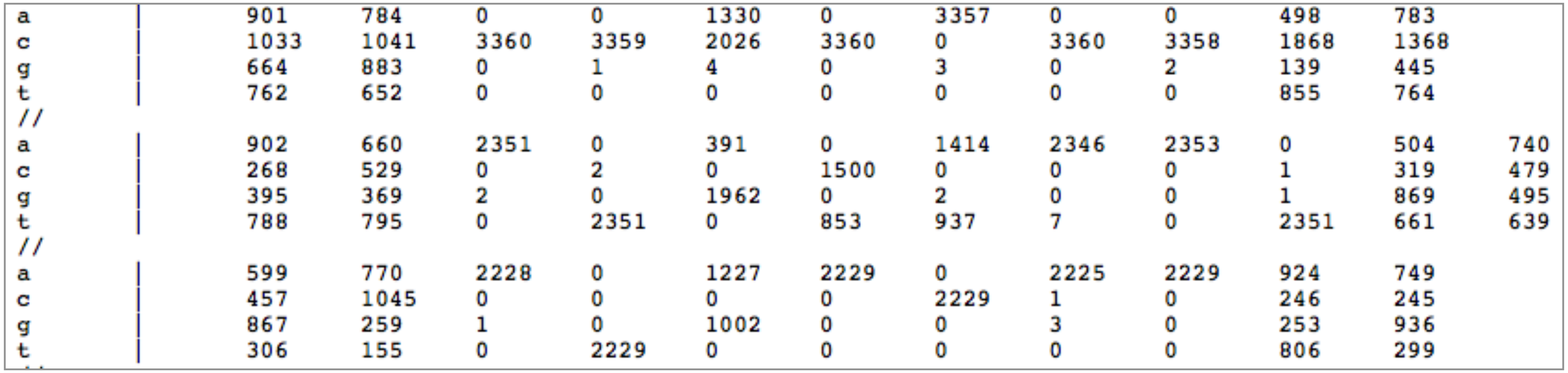
- The significance matrix can be used as “seed” to scan the input sequences and collect site.
- Those sites are in turn used to build a final matrix.

Significance matrix

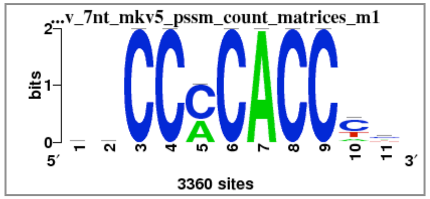


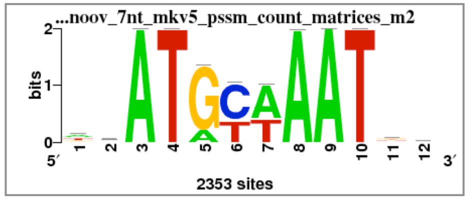
```
a     |     0     0     31.45 0     31.45 0     31.45 0     0     0     0
c     |     0     0     31.45 31.45 13.34 31.45 0     31.45 31.45 0     0
g     |     0     0     0     0     0     0     0     0     0     0     0
t     |     0     0     0     0     0     0     0     0     0     0     0
//
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
```

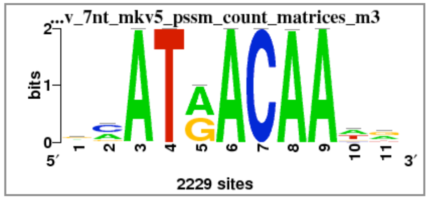
Final matrix



```
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
```

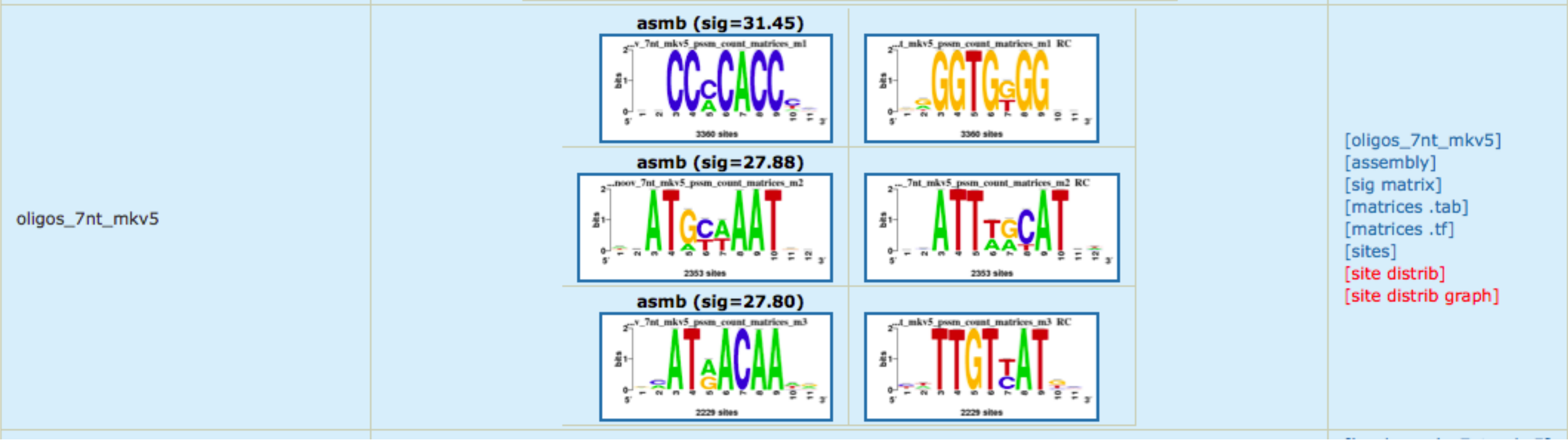






Motifs reported with oligo-analysis (Sox2 peaks from Chen, 2008)

- The program *oligo-analysis* detects over-represented words, as compared to some background model.
- For words of lenth *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

asmb (sig=31.45)

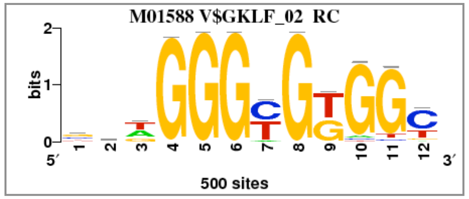
asmb (sig=27.88)

asmb (sig=27.80)

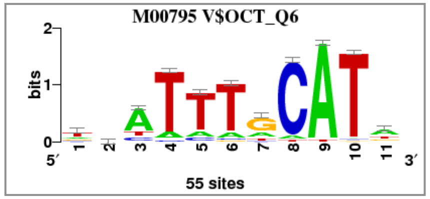
[oligos_7nt_mkv5]
[assembly]
[sig matrix]
[matrices .tab]
[matrices .tf]
[sites]
[site distrib]
[site distrib graph]

...v_7nt_mkv5_pssm_count_matrices_m1
3360 sites
...t_mkv5_pssm_count_matrices_m1 RC
3360 sites
...noov_7nt_mkv5_pssm_count_matrices_m2
2353 sites
..._7nt_mkv5_pssm_count_matrices_m2 RC
2353 sites
...v_7nt_mkv5_pssm_count_matrices_m3
2229 sites
...t_mkv5_pssm_count_matrices_m3 RC
2229 sites
bits

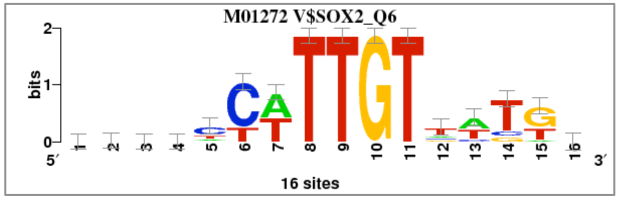
KLF (TRANSFAC built from Chen Klf4 set)



OCT (TRANSFAC, various OCT factors)

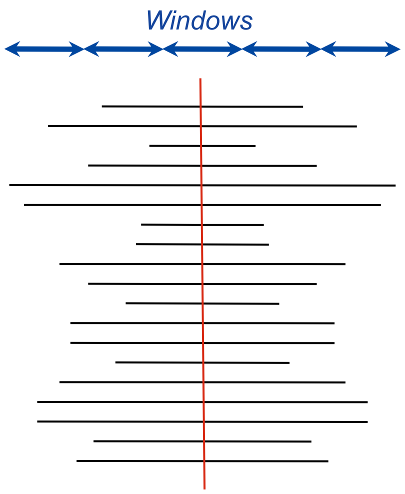


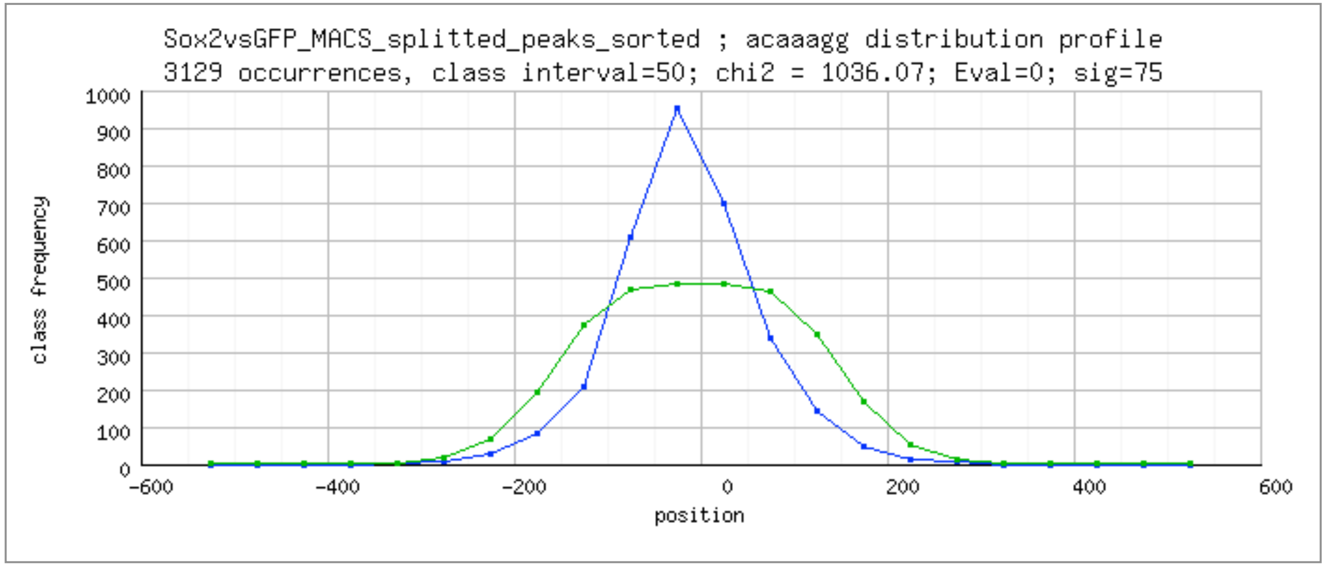
Sox2 (TRANSFAC, built from individual sites)



Detecting biases in word positions

- The program position-analysis (van Helden et al., 2000) detects words showing a heterogeneous distribution of occurrences across a set of input sequences.
- Principle: for each word
  - 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
  - Word length k=7
  - Reference position: the center of each peak.
  - The most significant word is ACAAAGG, which corresponds to the Sox2 consensus.



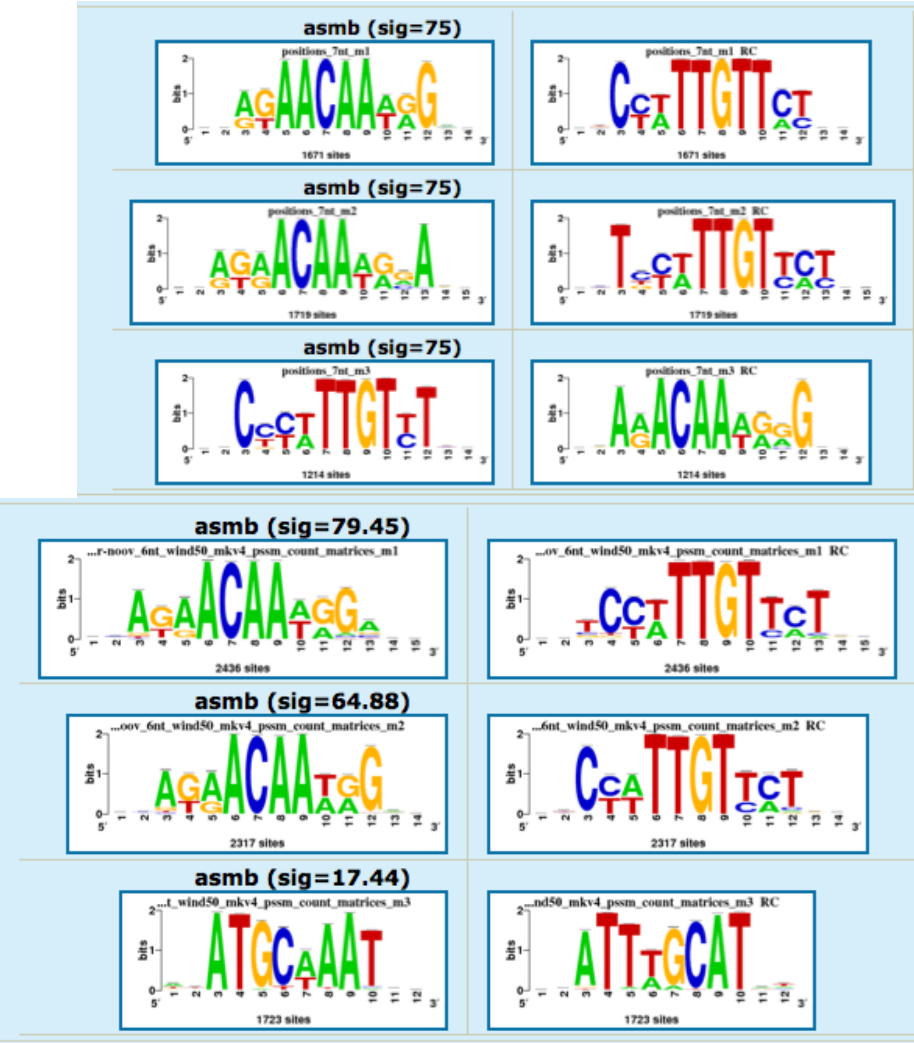


- Green: expected occurrences
  - Note: the expectation decreases with the distance to peak center because peaks have variable lengths.
- Blue: observed occurrences
  - The word ACAAGG is concentrated the center the ChIP-seq peak regions.

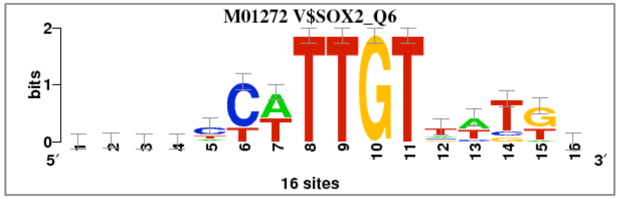
1. van Helden, J., del Olmo, M. and Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28, 1000-10.

Motifs with position biases in Sox2 peaks from Chen, 2008

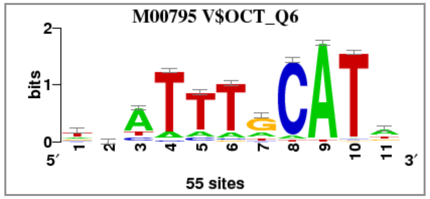
- *position-analysis*
  - detects the Sox2 motif in Sox2 peaks.
  - the partner motifs (Oct4, Klf4 are not detected).
- *local-words*
  - detects both the Sox2 and Oct4 motifs



Sox2 (TRANSFAC, built from individual sites)

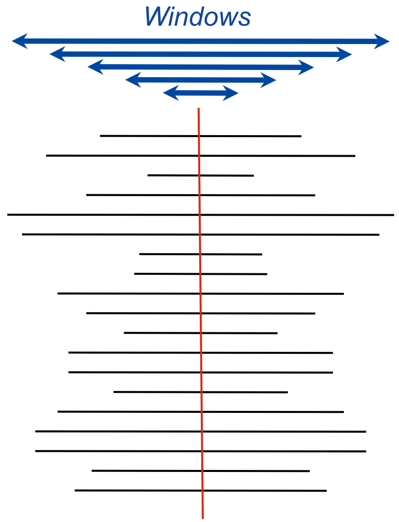


OCT (TRANSFAC, various OCT factors)



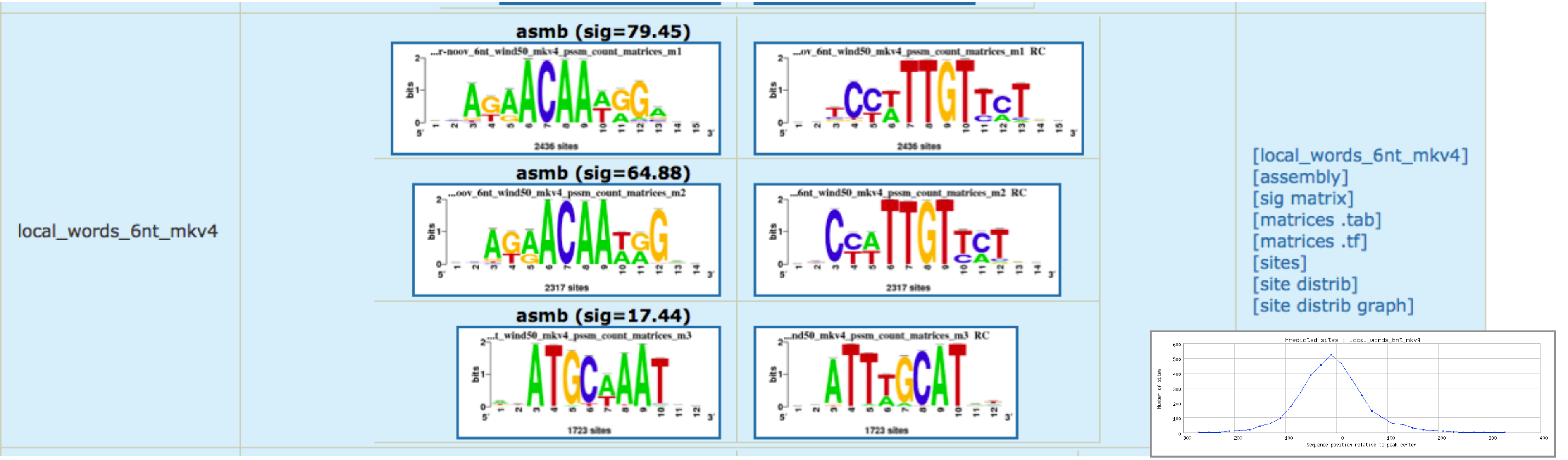
Local over-representation (program local-words)

- The program *local-words* detects words that are over-represented in specific position windows.
- The result is thus more informative than for *position-analysis*: in addition to the global positional bias, we detect the precise window where each word is over-represented.



Local over-representation (local-words)

- The program local-words detects windows of local over-representation.
- With windows of 50 bp, the program detects the Sox2 and Oct4 motifs.
- Those motifs are concentrated in the center of the peaks.



local_words_6nt_mkv4

asmb (sig=79.45)

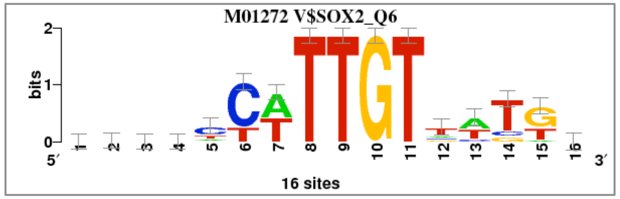
asmb (sig=64.88)

asmb (sig=17.44)

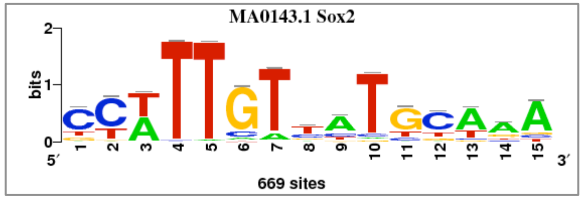
[local_words_6nt_mkv4]
[assembly]
[sig matrix]
[matrices .tab]
[matrices .tf]
[sites]
[site distrib]
[site distrib graph]

...r-noov_6nt_wind50_mkv4_pssm_count_matrices_m1
...ov_6nt_wind50_mkv4_pssm_count_matrices_m1 RC
2436 sites
...oov_6nt_wind50_mkv4_pssm_count_matrices_m2
...6nt_wind50_mkv4_pssm_count_matrices_m2 RC
2317 sites
...t_wind50_mkv4_pssm_count_matrices_m3
...nd50_mkv4_pssm_count_matrices_m3 RC
1723 sites
bits
Predicted sites : local_words_6nt_mkv4
Number of sites
Sequence position relative to peak center

Sox2 (TRANSFAC, built from individual sites)

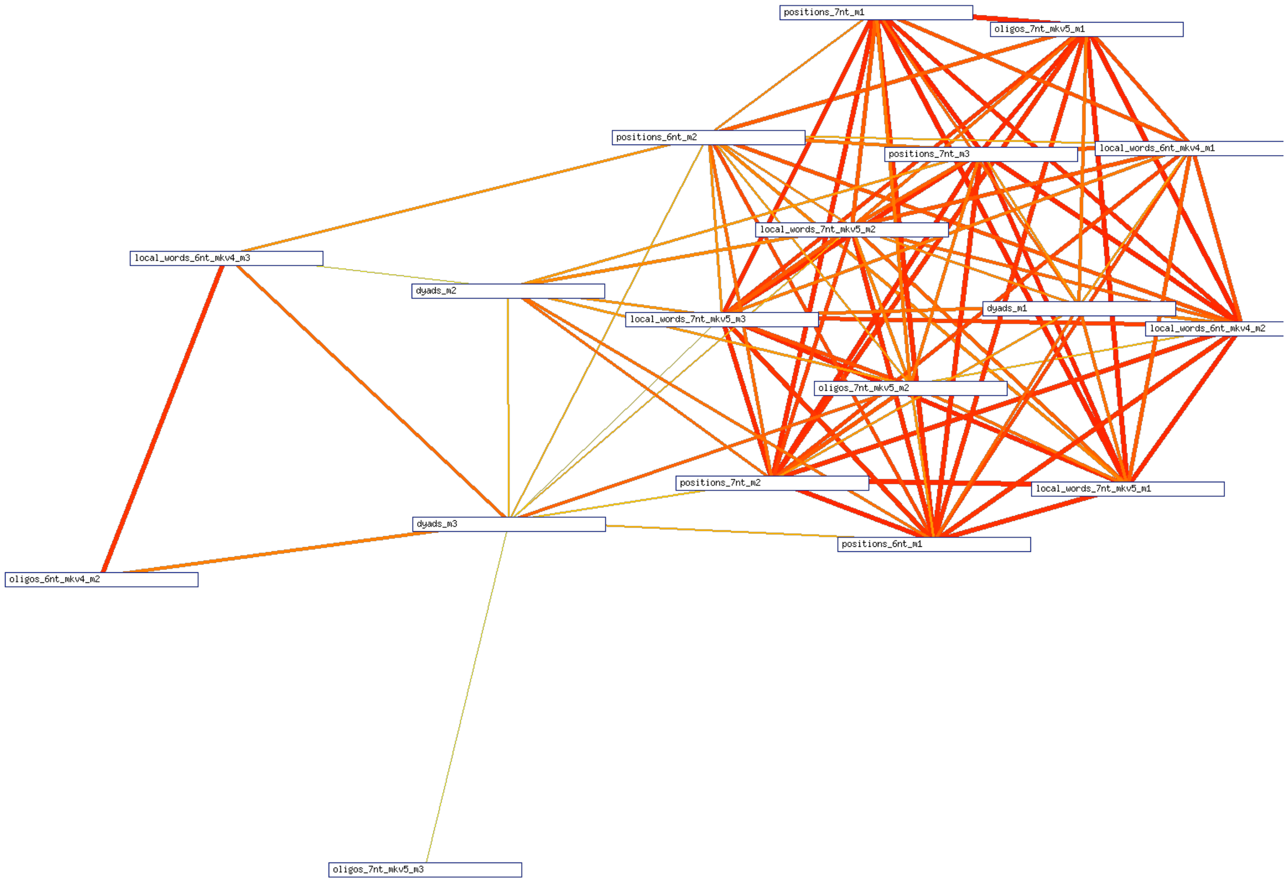


SOCT (JASPAR, built from Chen Sox2 set)



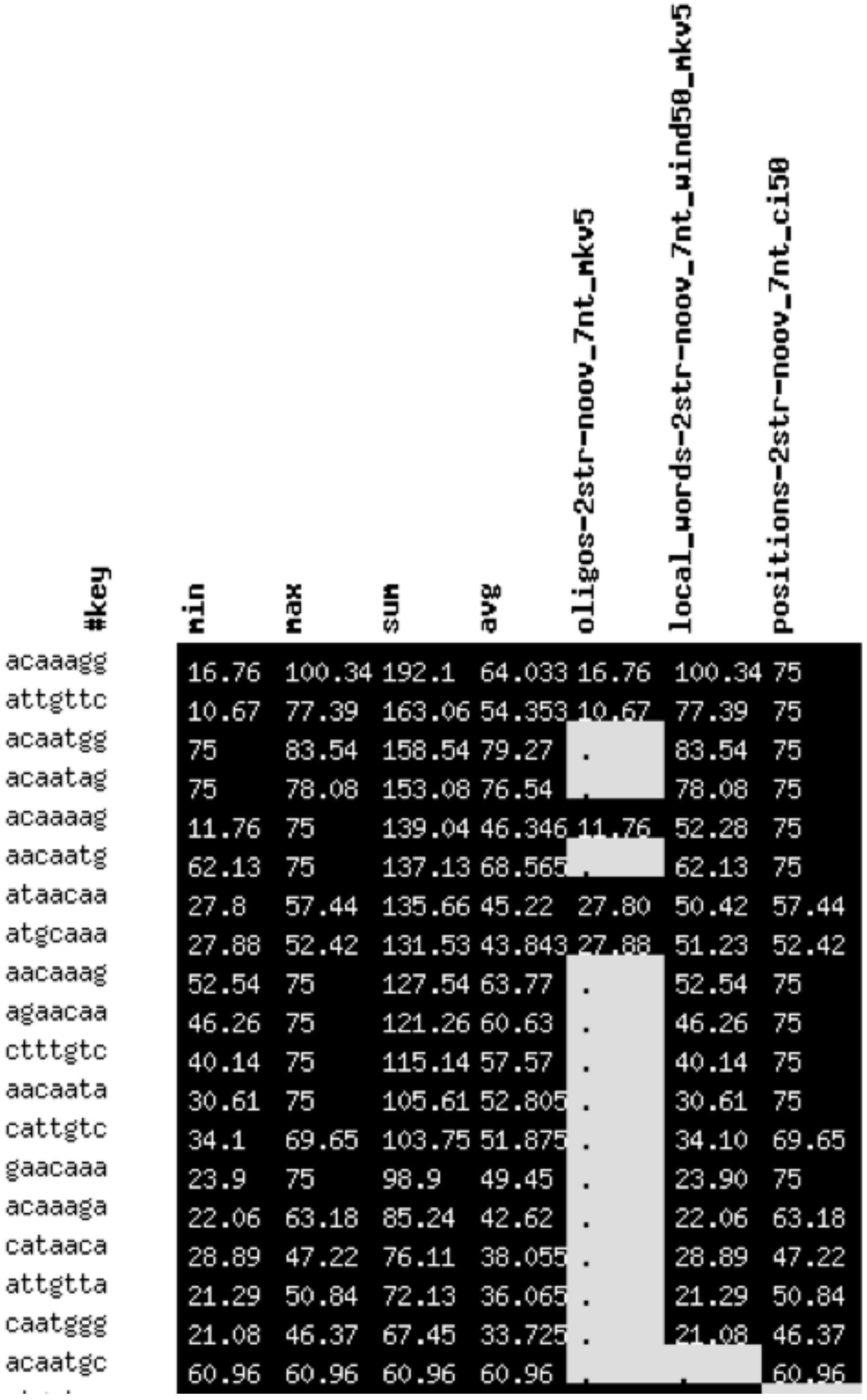
Comparisons between discovered motifs

- Pairwise comparisons show the consistency between the motifs discovered by the different approaches.



Word merging

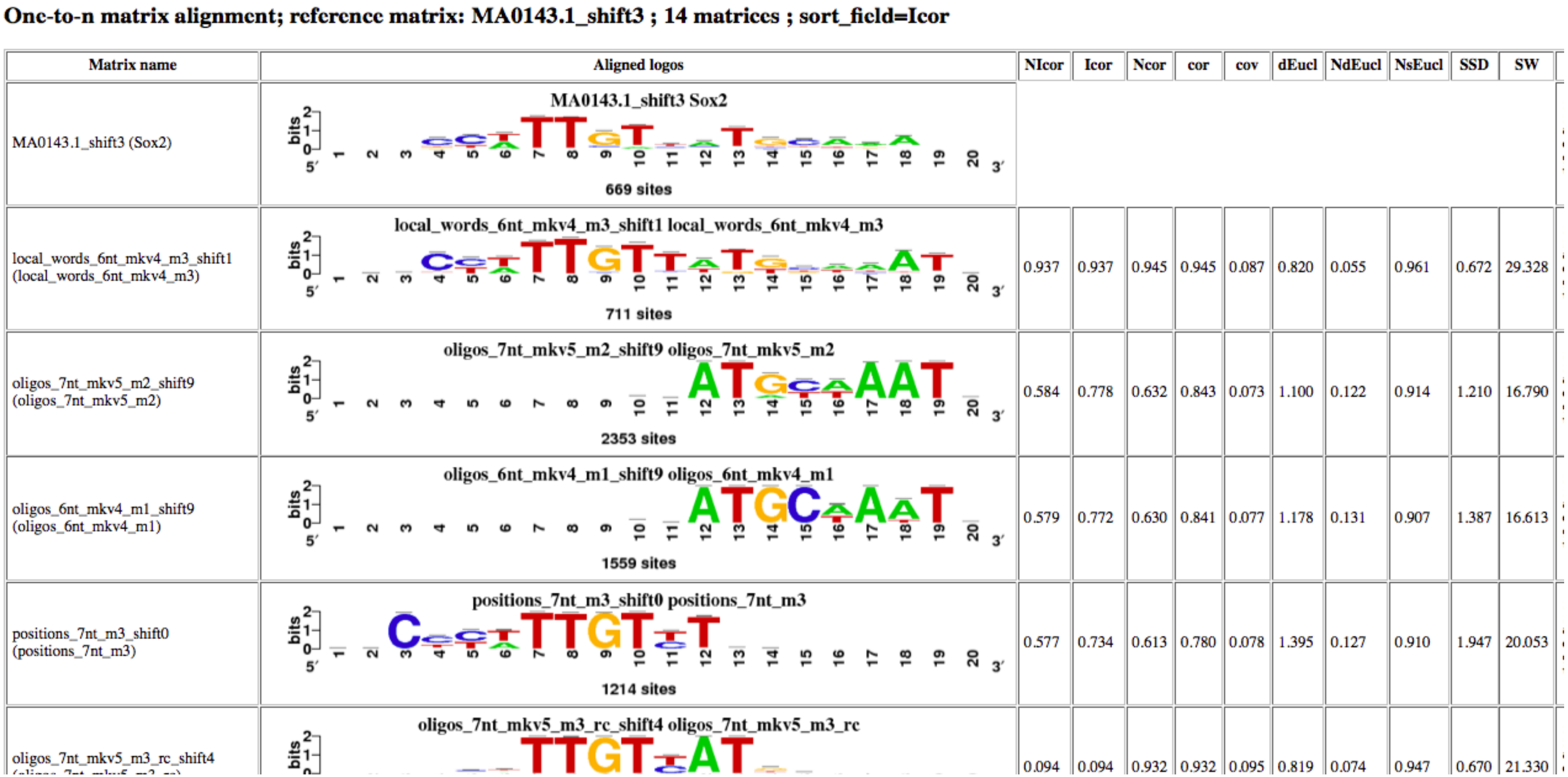
- The words discovered by the different approaches can be compared and merged into a word significance table.
- The most significant and consistent words (discovered by several approaches) are used as seeds to collect final matrices.



#key	min	max	sum	avg	oligos-2str-noov_7nt_mkv5	local_words-2str-noov_7nt_wind50_mkv5	positions-2str-noov_7nt_ci50
acaaagg	16.76	100.34	192.1	64.033	16.76	100.34	75
attgttc	10.67	77.39	163.06	54.353	10.67	77.39	75
acaatgg	75	83.54	158.54	79.27	.	83.54	75
acaatag	75	78.08	153.08	76.54	.	78.08	75
acaaaag	11.76	75	139.04	46.346	11.76	52.28	75
aacaatg	62.13	75	137.13	68.565	.	62.13	75
ataacaa	27.8	57.44	135.66	45.22	27.80	50.42	57.44
atgcaaa	27.88	52.42	131.53	43.843	27.88	51.23	52.42
aacaaag	52.54	75	127.54	63.77	.	52.54	75
agaacaa	46.26	75	121.26	60.63	.	46.26	75
ctttgtc	40.14	75	115.14	57.57	.	40.14	75
aacaata	30.61	75	105.61	52.805	.	30.61	75
cattgtc	34.1	69.65	103.75	51.875	.	34.10	69.65
gaacaaa	23.9	75	98.9	49.45	.	23.90	75
acaaaga	22.06	63.18	85.24	42.62	.	22.06	63.18
cataaca	28.89	47.22	76.11	38.055	.	28.89	47.22
attgtta	21.29	50.84	72.13	36.065	.	21.29	50.84
caatggg	21.08	46.37	67.45	33.725	.	21.08	46.37
acaatgc	60.96	60.96	60.96	60.96	.	.	60.96

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, ...

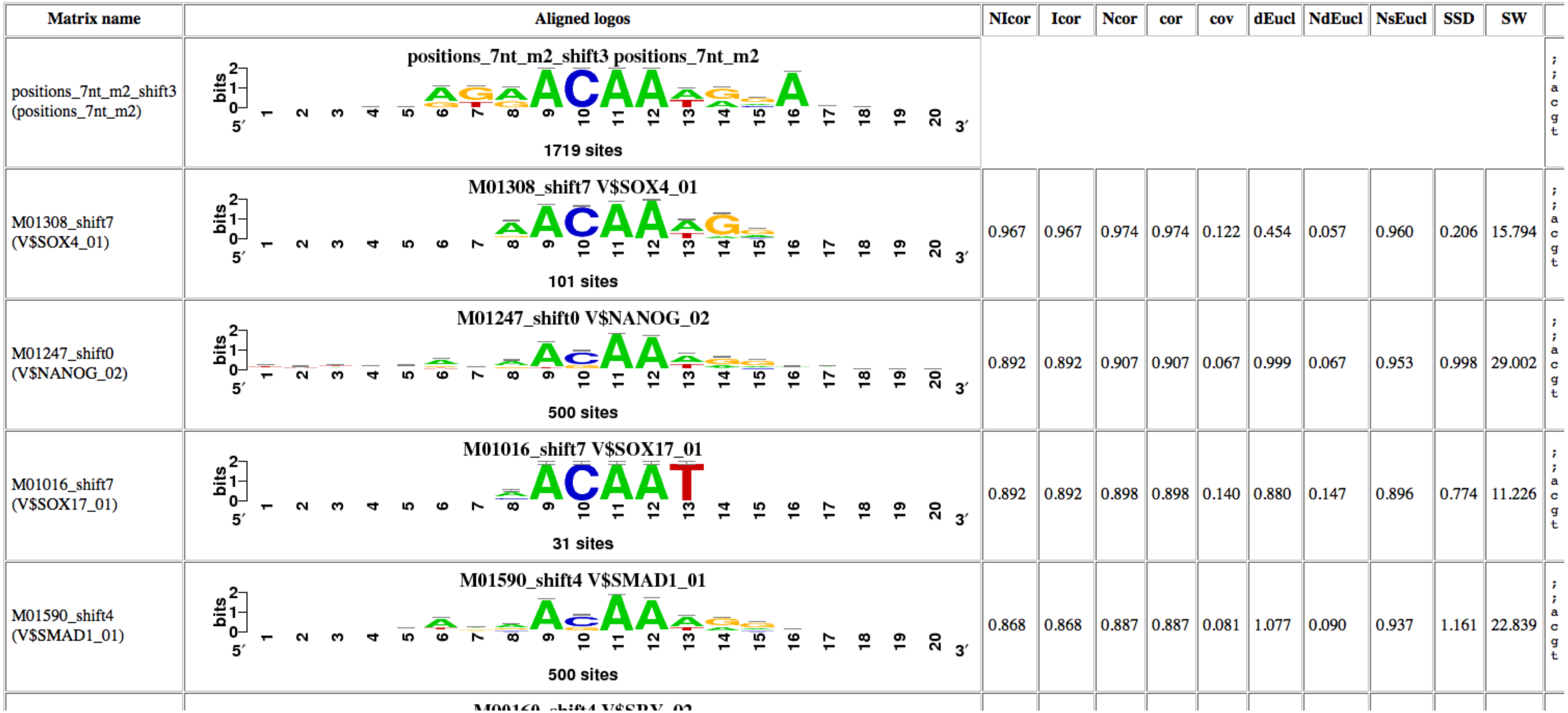


One-to-n matrix alignment; reference matrix: MA0143.1_shift3 ; 14 matrices ; sort_field=Icor

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 motifs versus databases (TRANSFAC, JASPAR, ...)

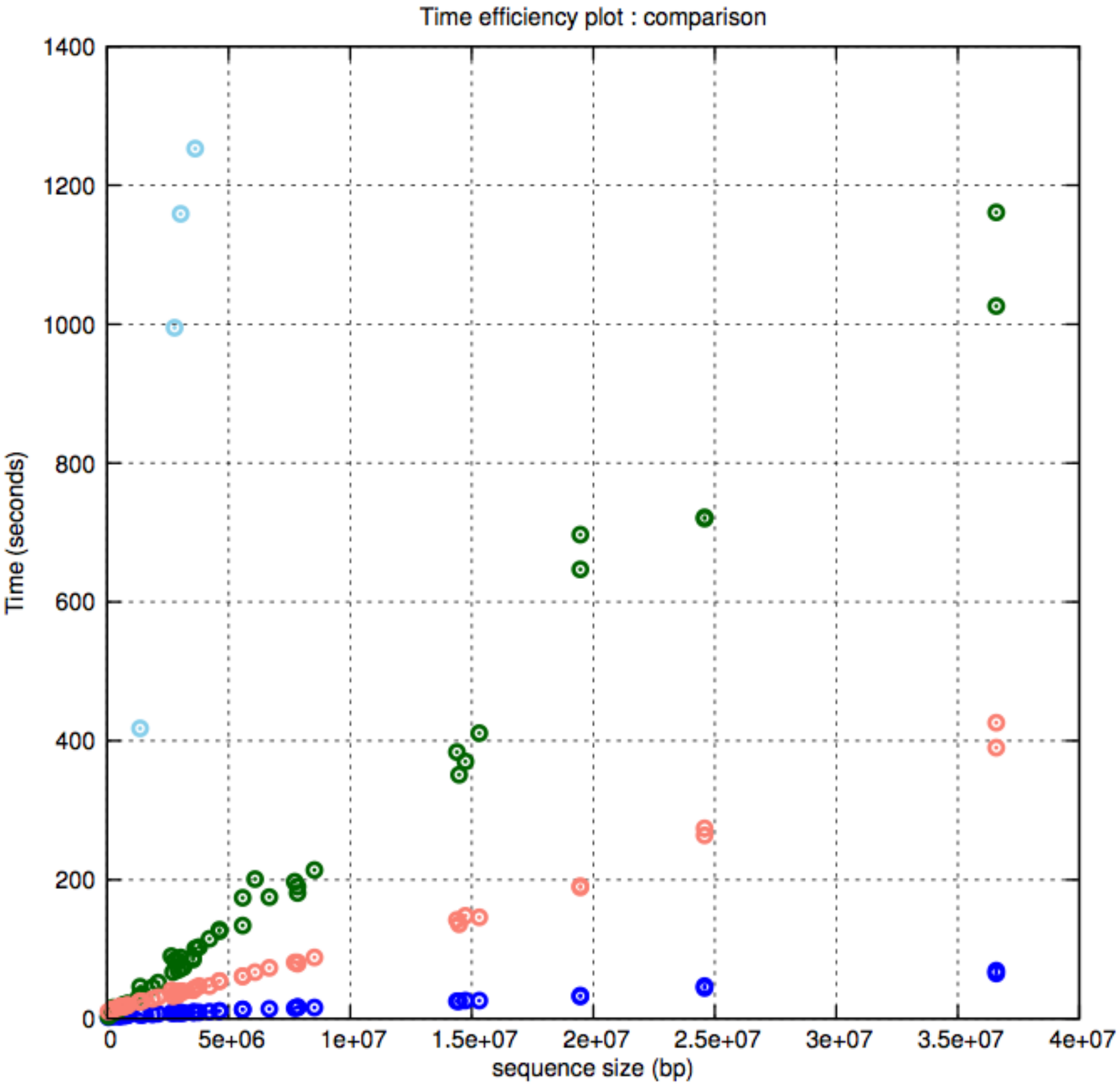
- Discovered motifs are compared to all motifs



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											

Time efficiency : position-analysis

- 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.
- On my laptop (MacBook Pro, 8Gb RAM), the biggest files (37Mb) are treated in
  - 69 seconds with *oligo-analysis*
  - 7 minutes with *dyad-analysis*
  - 20 minutes with *position-analysis*



Conclusions

Conclusions

- The program ***peak-motifs*** provides a flexible tool for analyzing motifs in large collections of peaks.
  - Time-linear algorithms.
  - Reduced memory usage.
- The work flow provides an integrated view of all steps from peaks to motifs.
  - Sequence length distribution
  - Composition analysis
  - Motif discovery
  - Positional distribution of the discovered motifs
  - Comparison of discovered motifs with
    - reference motifs
    - motif databases
- Web interface
  - Simplcity of use (“one click” interface).
  - Advanced options can be accessed optionally.
  - Allows to analyze data set of realistic size (uploaded files).
- Perspectives
  - Predicting the most likely site sinside the peaks.
  - Interfacing the results with genome browsers (UCSC) for direct visualization of the predicted sites.
  - Integrating additional motif discovery software (MEME, info-gibbs) to evaluate the robustness of the motifs.