Regulatory Sequence Analysis

RSAT peak-motifs

Fast extraction of transcription factor binding motifs from full-size ChIP-seq datasets

Jacques van Helden

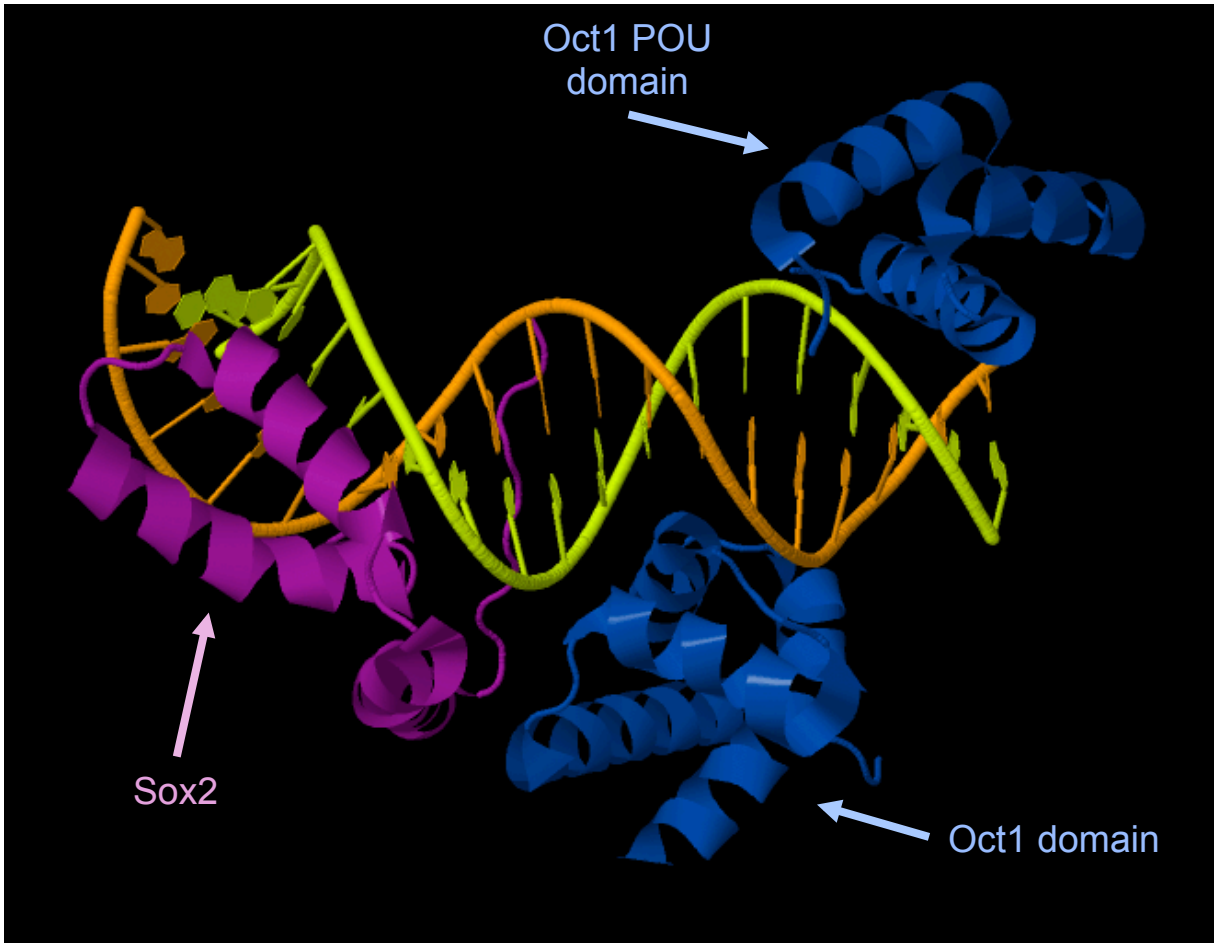
Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)

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

Sox2/Oct4 cooperative binding

- The Sox2 and Oct 4 transcription factors recognize specific DNA motifs.
- Cooperative binding: Sox2 and Oct4 closely interact to bind DNA.
- The pair of transcription factors recognizes a composite motif called the « SOCT » motif (SOx+OCT).



http://www.pdb.org/pdb/explore/explore.do?structureId=1O4X

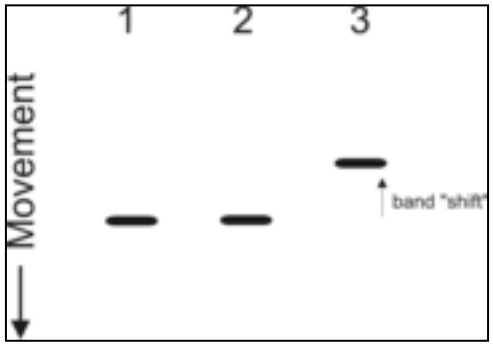
Transcription factor binding site prediction : difficulties

- 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.

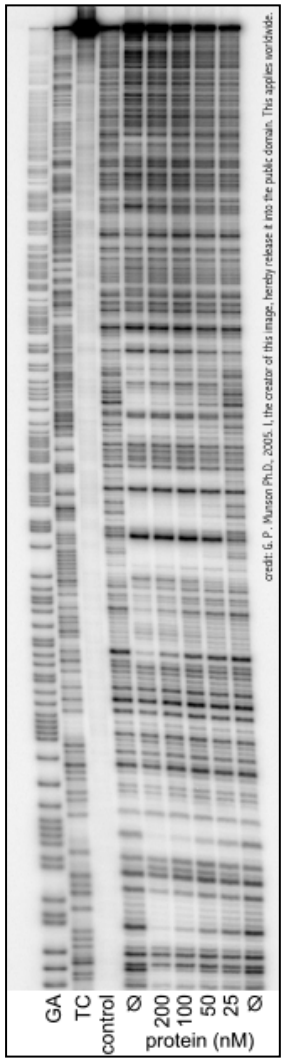
- Transcription factor binding motifs are poorly informative.
  - Motif width varies from 5 to 25 base pairs (some factors bind spaced motifs).
  - Typically 5-10 partly conserved positions.
  - Predicting individual binding sites at a genome scale is expected to return many false positives.

- 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



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(3): 808-824.

Sox2 : from binding sites to binding motif

- The TRANSFAC database contains collections of experimentally proven binding sites for several hundreds transcription factors.
- Those binding sites can be used to build motifs, that represent the specificity of the transcription factor.

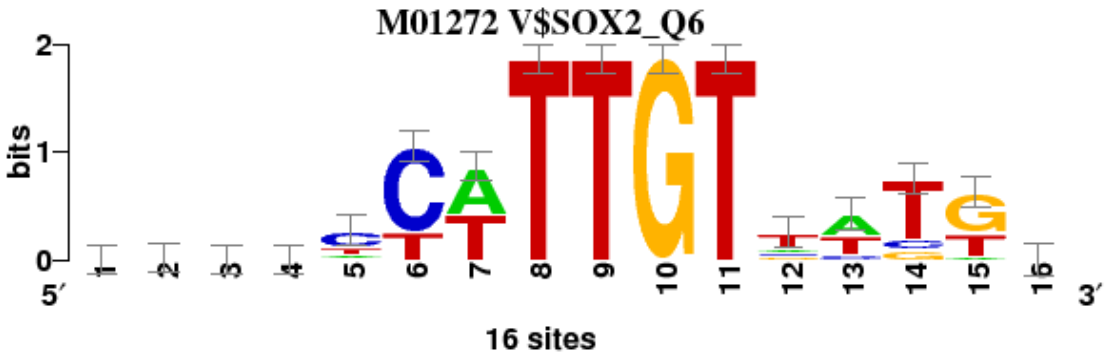
Collection of binding sites
used to build the Sox2 matrix
(TRANSFAC M01272)

```
R15133   GCCCTCATTGTTATGC
R15201   AAACTCTTTGTTTGGA
R15231   TTCACCATTGTTCTAG
R15267   GACTCTATTGTCTCTG
R16367   GATATCTTTGTTTCTT
R17099   TGCACCTTTGTTATGC
R19276   AATTCCATTGTTATGA
R19367   AAACTCTTTGTTTGGA
R19510   ATGGACATTGTAATGC
R22342   AGGCCTTTTGTCCTGG
R22344   TGTGCTTTTGTNNNNN
R22359   CTCAACTTTGTAATTT
R22961   GCAGCCATTGTGATGC
R23679   CACCCCTTTGTTATGC
R25928   TTTTCTATTGTTTTTA
R27428   AAAGGCATTGTGTTTC
```

Position-specific scoring matrix (PSSM)

A	6	7	4	4	2	0	**8**	0	0	0	0	2	**7**	0	1	4
C	2	2	6	5	**9**	**12**	0	0	0	0	0	2	2	2	0	6
G	4	3	2	4	1	0	0	0	0	**16**	0	2	0	2	**9**	3
T	4	4	4	3	**4**	**4**	**8**	**16**	**16**	0	**16**	**9**	**6**	**11**	**5**	2

Sequence logo



"Family" motifs

- TRANSFAC contains a matrix representing the "consensus" of the binding specificity for several transcription factors belonging to the OCT family.
- This matrix was built from 55 sites, collected from different organisms (mouse, human, cat, xenopus, ...).

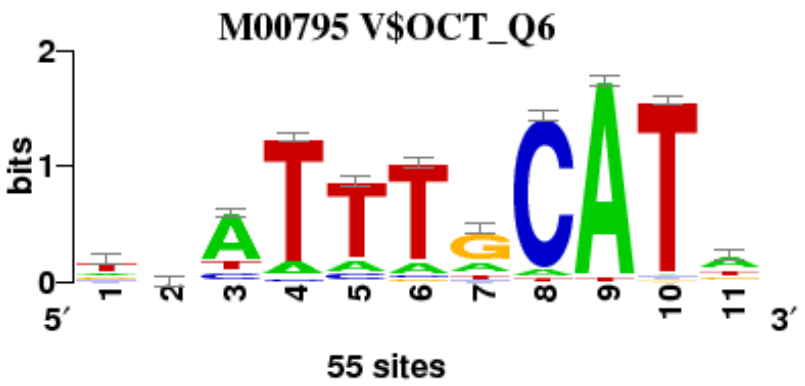
Collection of binding sites
used to build the motif of the OCT ***family*** *(TRANSFAC M00795)*

```
R00306   TAATTAGCATA
R00551   ATATTTGCATT
R00662   TTATTTGCATA
R00664   TCATTTGCATA
R00666   ACATTTGCATA
R00814   TCGTTAGCATG
R00815   CGCATGGCATC
R00820   GGAATTCCATT
R00824   CGTATCTCATT
R00834   TTATTTGCATA
R00842   GGATTTGCATA
R00855   GTATTTGCATA
R00872   TAATTTGCATT
R00888   CGATTTGCATA
R00893   TGATTTGCATA
... 40 other sites
```

Position-specific scoring matrix (PSSM)

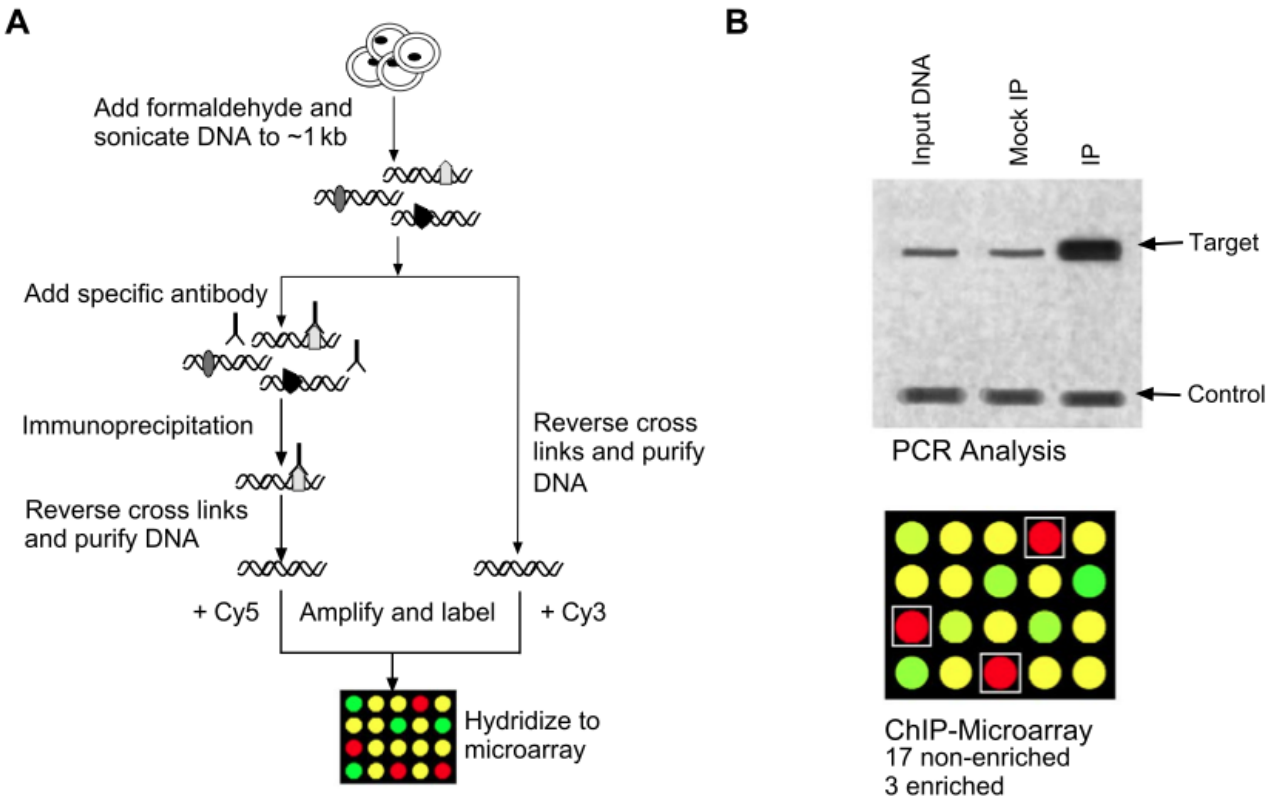
A	10	14	**37**	6	7	6	**11**	3	**53**	1	**27**
C	7	12	7	2	5	2	3	**50**	0	1	4
G	10	15	2	0	1	2	**34**	0	0	1	10
T	**28**	14	9	**47**	**42**	**45**	7	2	2	**52**	14

Sequence logo



ChIP-on-chip

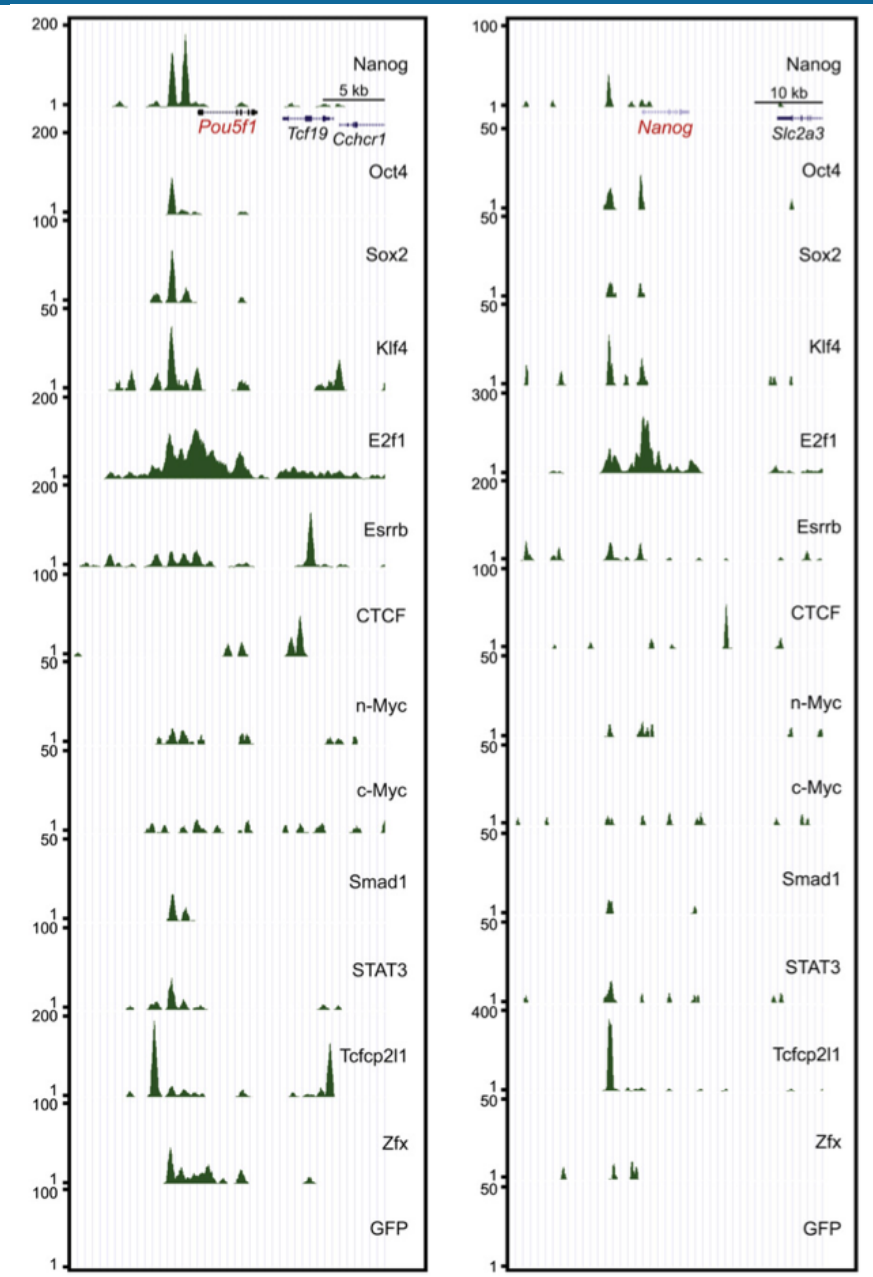
- The ChIP-on-chip method combines
  - Chromatin Immunoprecipitation (ChIP) to select genome fragments bound to a tagged transcription factor.
  - DNA microarrays (chip) spotted with several thousands of genome fragments (typically all the intergenic regions of agiven organism) are used to detect the 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.
  - Next Generation Sequencing (NGS) to characterize the immunoprecipitated DNA fragments.
- 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 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).

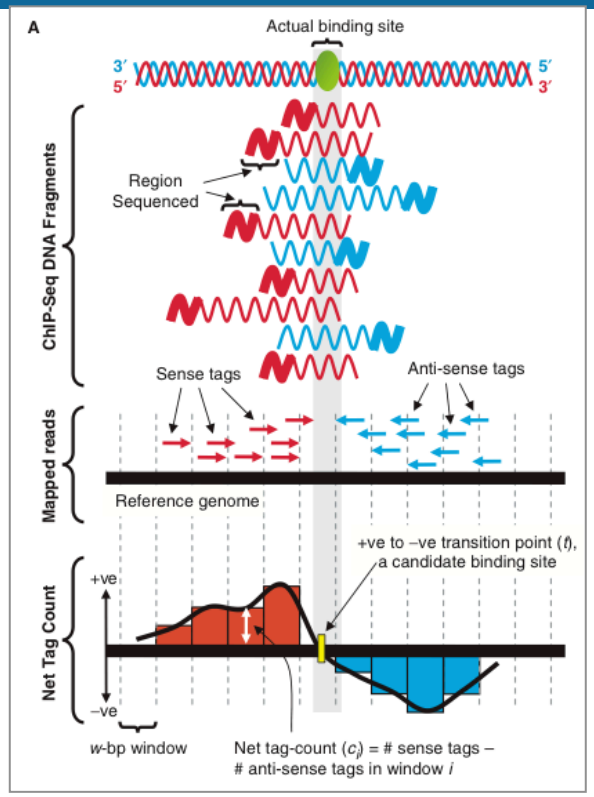


Figure from Jothi et al. (2008)

- Figure
  - RMP: read per millions.

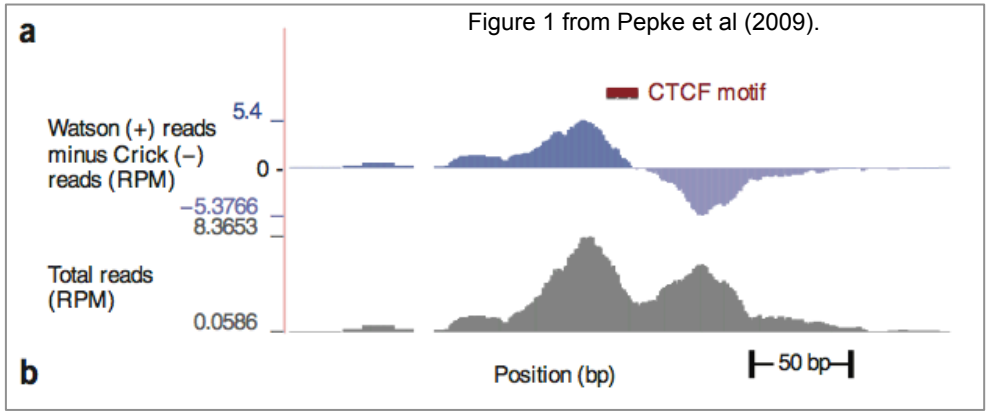
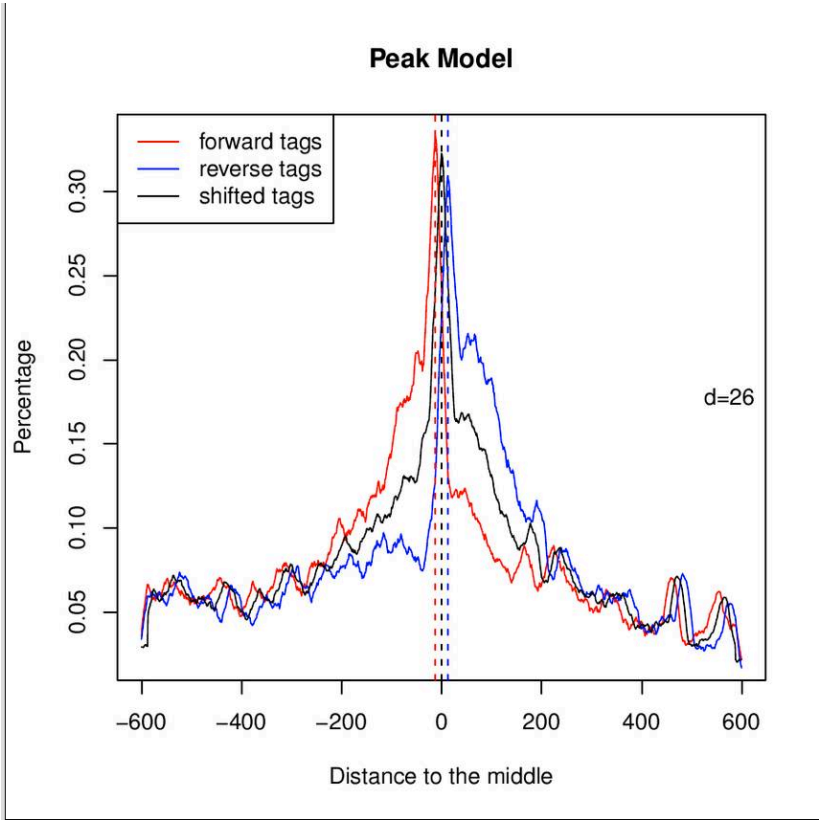


Figure 1 from Pepke et al (2009).

- 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 (http://main.g2.bx.psu.edu/)
- The curves indicate the density 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 in this case.



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.

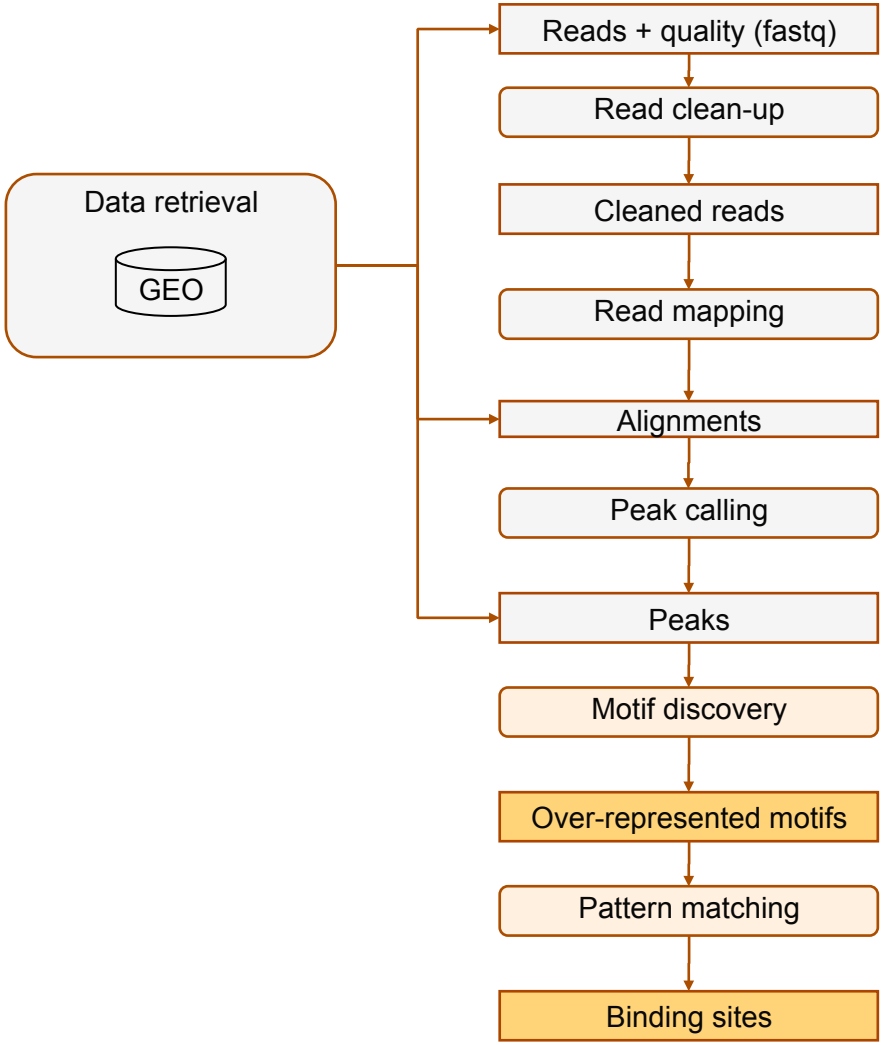
Challenges

- 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

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***: chromosomal coordinates of each read.
  - ***Peak locations***: several thousands of variable size regions (typically between 100bp and 10kb).
- A technological bottleneck lies in the next step: exploitation of full peak collections to discover motifs and predict binding sites.



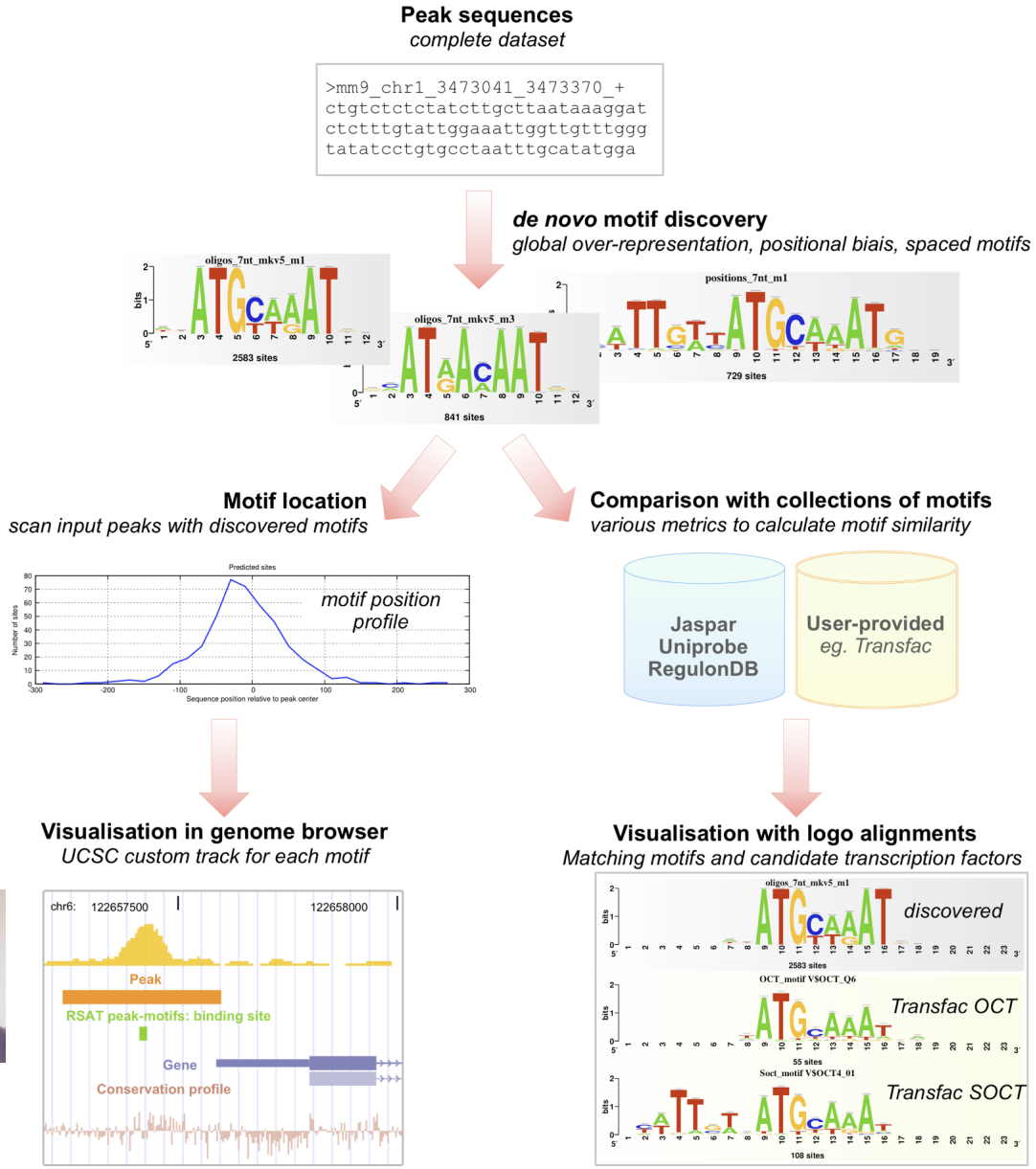
Software tools for analyzing motifs in ChIP-seq peak sets

Program	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs	RSAT *peak-motifs*
Web interface	**yes**	**yes**	**yes**	no	no	**yes**
Size limitation	100kb (web site)	500kb (web site)	unrestricted, but motif discovery restricted to 600 peaks clipped to 100bp	motif discovery restricted to a few hundred base pairs	-	unrestricted (Web site tested with 22 Mb)
Stand-alone version	**yes**	no	**yes**	**yes**	**yes**	**yes**
Tasks						
peak finding	no	no	no	**yes**	no	no
annotation of peak-flanking genes	no	**yes**	no		no	no
sequence composition (mono- and di-nucleotides)	no	no	no		no	**yes**
motif discovery	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
enrichment in motifs from databases	no	**yes**	**yes**		no	no
enrichment in discovered motifs	no	no	no		no	**yes**
peak scoring	no	no	**yes**	**yes**	no	no
motif clustering	no	no	no		**yes**	no
comparison discovered motifs / motif DB	no	no	**yes**		**yes**	**yes**
sequence scanning for site prediction	no	no	**yes**		no	**yes**
positional distribution of sites inside peaks	no	**yes**	no		**yes**	**yes**
visualization in genome browsers	no	**yes**	no		no	**yes**
Motif discovery algorithms	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk

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.

RSAT peak-motifs: specialized work flow for motif analysis in ChIP-seq peaks

- The program ***peak-motifs*** is a work flow combining a series of RSAT tools optimized discovered motifs in large sequence sets (te 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)

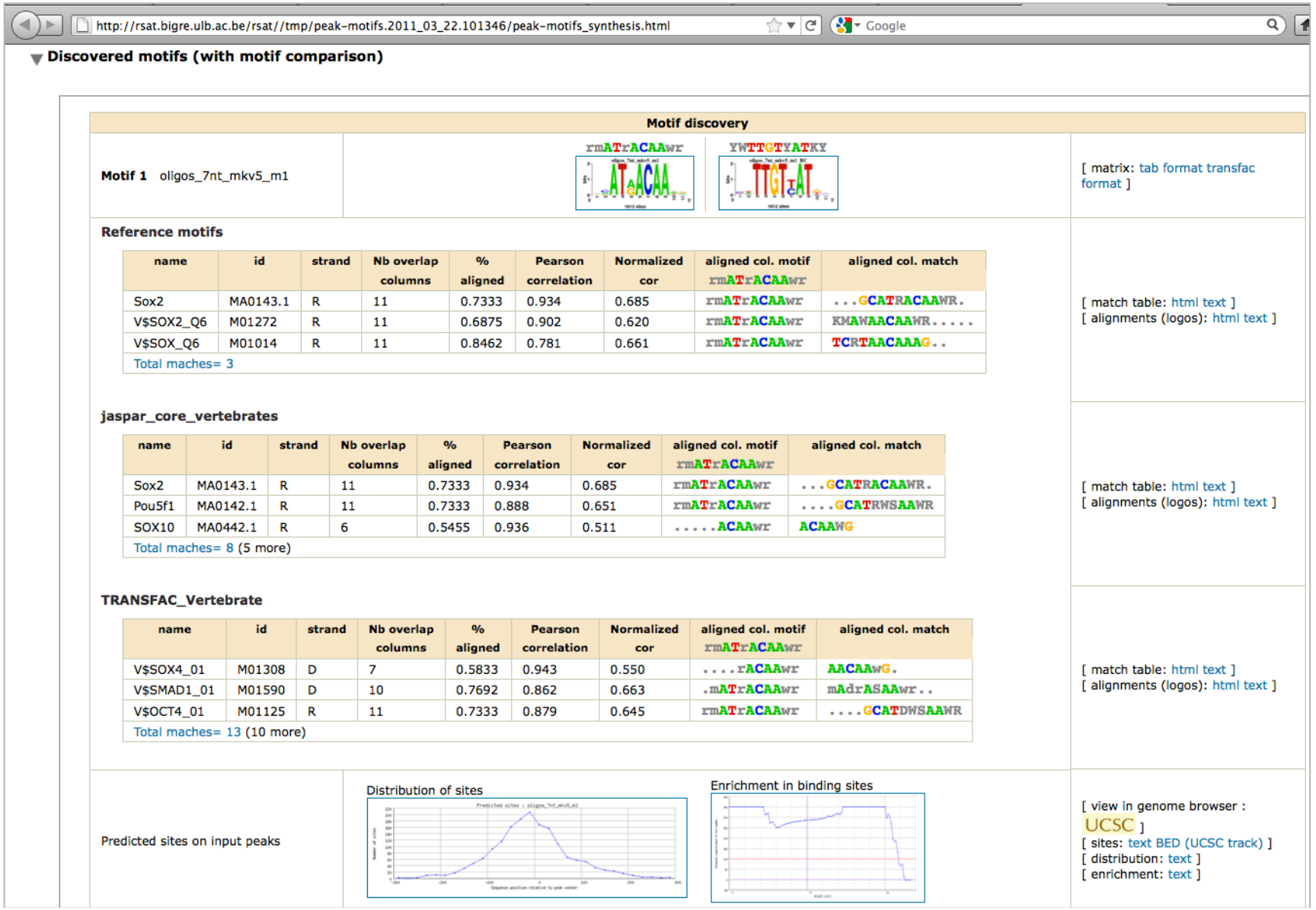


Morgane Thomas-Chollier | Matthieu Defrance | Olivier Sand | Carl Herrmann (TACG, Marseille) | Denis Thieffry ENS Paris

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.

Report for a discovered motif

- The result is reported as a HTML page with one summary per **discovered motif** + correspondences to **known TF motifs** + links to all the detailed result files + feature loading in **UCSC genome browser**.



http://rsat.bigre.ulb.ac.be/rsat//tmp/peak-motifs.2011_03_22.101346/peak-motifs_synthesis.html

Discovered motifs (with motif comparison)

Motif discovery

Motif 1 oligos_7nt_mkv5_m1

rmATrACAAwr YWTTGTYATKY

[matrix: tab format transfac format]

Reference motifs

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rmATrACAAwr	aligned col. match
Sox2	MA0143.1	R	11	0.7333	0.934	0.685	rmATrACAAwr	...GCATRACAAWR.
V$SOX2_Q6	M01272	R	11	0.6875	0.902	0.620	rmATrACAAwr	KMAWAACAAWR.....
V$SOX_Q6	M01014	R	11	0.8462	0.781	0.661	rmATrACAAwr	TCRTAACAAAG..

Total maches= 3

[match table: html text]
[alignments (logos): html text]

jaspar_core_vertebrates

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rmATrACAAwr	aligned col. match
Sox2	MA0143.1	R	11	0.7333	0.934	0.685	rmATrACAAwr	...GCATRACAAWR.
Pou5f1	MA0142.1	R	11	0.7333	0.888	0.651	rmATrACAAwr	....GCATRWSAAWR
SOX10	MA0442.1	R	6	0.5455	0.936	0.511	.....ACAAwr	ACAAWG

Total maches= 8 (5 more)

[match table: html text]
[alignments (logos): html text]

TRANSFAC_Vertebrate

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rmATrACAAwr	aligned col. match
V$SOX4_01	M01308	D	7	0.5833	0.943	0.550	....rACAAwr	AACAAwG.
V$SMAD1_01	M01590	D	10	0.7692	0.862	0.663	.mATrACAAwr	mAdrASAAwr..
V$OCT4_01	M01125	R	11	0.7333	0.879	0.645	rmATrACAAwr	....GCATDWSAAWR

Total maches= 13 (10 more)

[match table: html text]
[alignments (logos): html text]

Predicted sites on input peaks

Distribution of sites

Enrichment in binding sites

[view in genome browser : UCSC]
[sites: text BED (UCSC track)]
[distribution: text]
[enrichment: text]

Case study 1: Embryonic Stem (ES) cell pluripotency and self-renewal

Embryonic Stem (ES) cell pluripotency and self-renewal

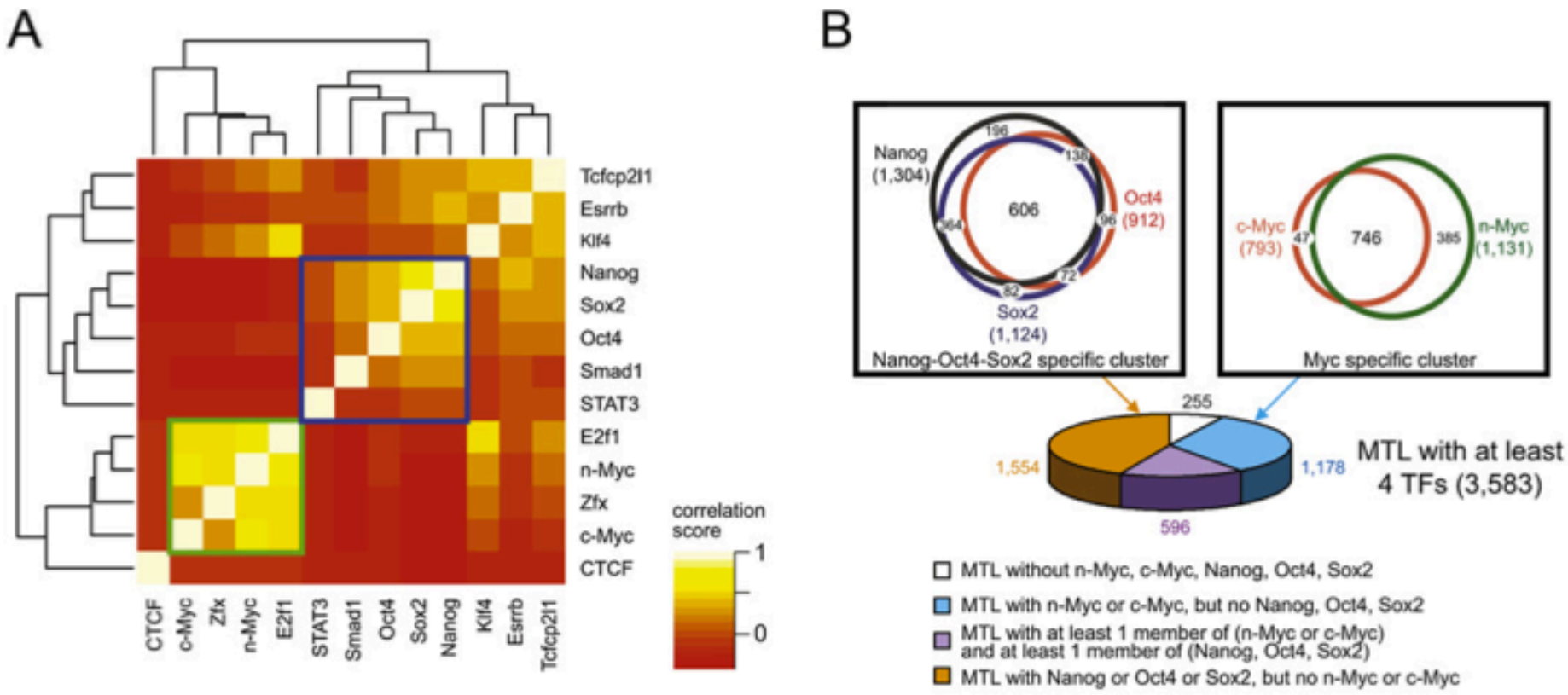
- Chen et al. (2008) used ChIP-seq to characterize the binding profiles of 12 factors involved in ES cell pluripotency and self-renewal.
- For some of those factors, known motifs can be found in databases (TRANSFAC, Jaspar).
- We selected those annotated motifs as reference for assessing pattern discovery programs.
- We preferably selected motifs built from dedicated experiments (footprints, EMSA, SELEX).
- For some factors the only matrices available from motif discovery in ChIP-seq peaks.

TF	**Role in Embryonic Stem (ES) cells**
Nanog	pluripotency, self-renewal
Oct4	pluripotency, self-renewal
Sox2	pluripotency, self-renewal
Esrrb	pluripotency, self-renewal
Zfx	pluripotency, self-renewal
Smad1	BPM signaling pathway component
STAT3	LIF signaling pathway component
Tcfcp2l1	preferentially upregulated in ES cells; unkown function
E2F1	regulating cell-cycle progression
CTCF	transcriptional insulation
Klf4	Fibroblast reprogramming to induced pluripotent stem cells
Klf4	Fibroblast reprogramming to induced pluripotent stem cells
c-Myc	Fibroblast reprogramming to induced pluripotent stem cells
n-Myc	

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

Multiple-factor binding loci

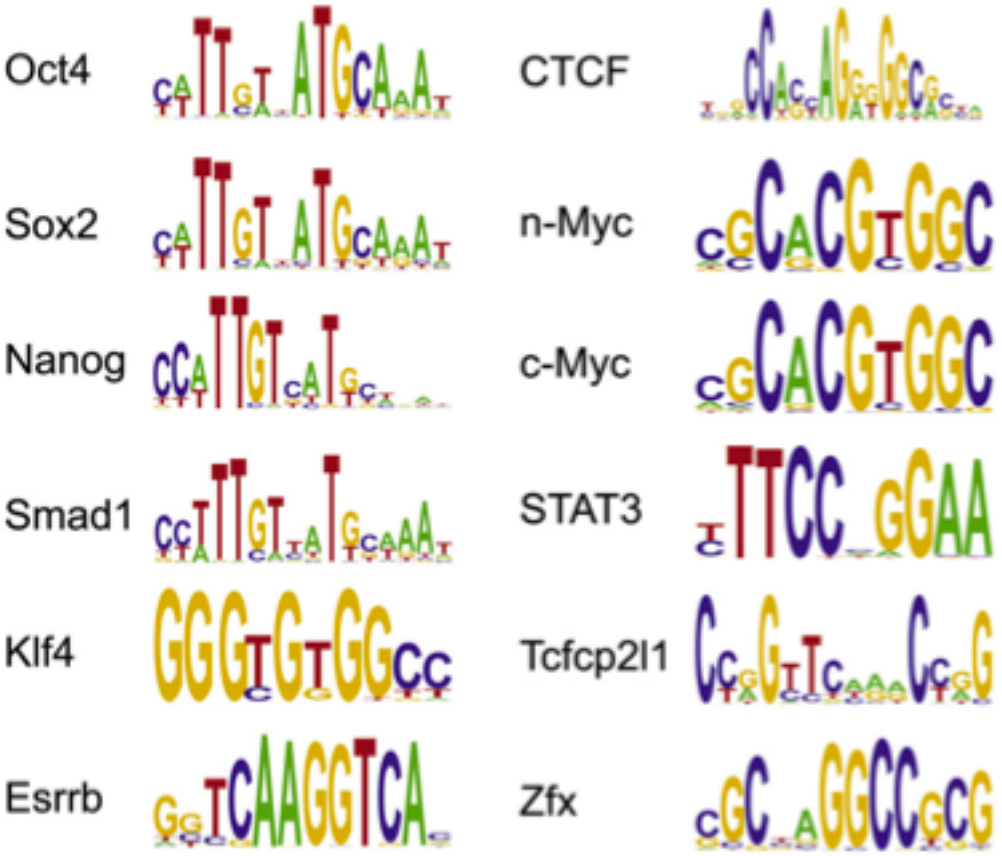
- Chen et al (2008) show that some groups of transcription factors show a strong tendency to co-occur in the same loci.
  - Cluster 1: Sox2 + Oxt4 + Nanog
  - Cluster 2: c-Myc + n-Myc



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

Case study 1: Chen et al. 2008

- Chen et al (2010) characterized the binding location of 13 mouse transcription factors involved in embryonic stem cell pluripotency and self-renewal.
- Combined the motif discovery tools Weeder and NMICA to predict motifs in each set of ChIP-seq peaks.
  - Motif prediction was restricted to the 500 top-scoring peaks.
  - Several data sets reveal the same composite motif: the SOCT motif indicating the Sox2 / Oct4 cooperative binding is found in Oct4, Sox2, Nanog and Smad1 peaks.

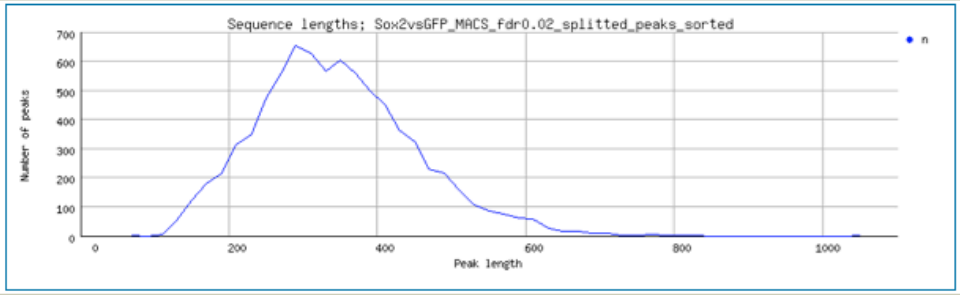
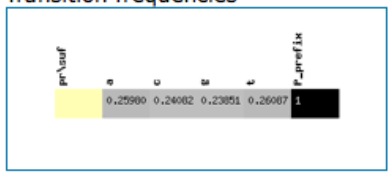
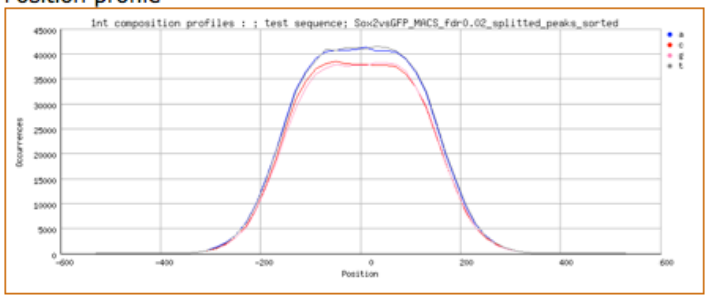
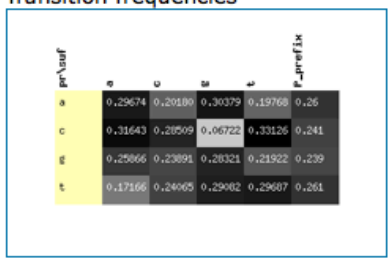
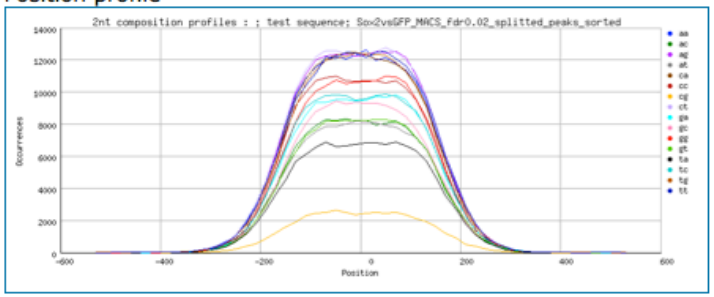


Chen et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell (2008) vol. 133 (6) pp. 1106-17

Composition analysis

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

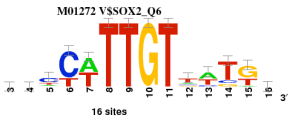
Sequence composition (peak sequences)

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  Position profile 	[1nt: freq transitions] [bg model: Inclusive format] [profile: table html(individual)]
2nt composition	Transition frequencies  Position profile 	[2nt: freq transitions] [bg model: Inclusive format] [profile: table html(individual)]

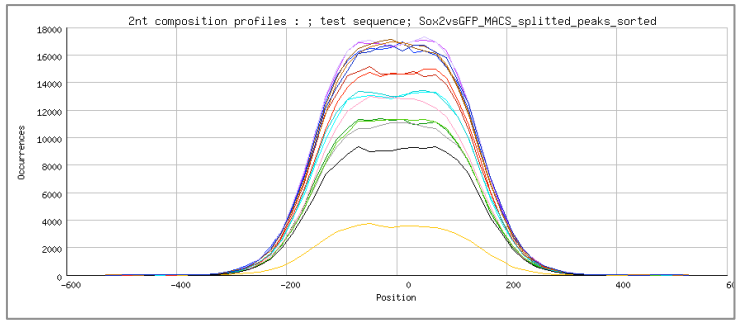
Composition analysis results

- The composition analysis reveals differences between data sets.
  - Sox2 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.

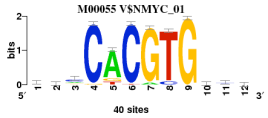
Sox2



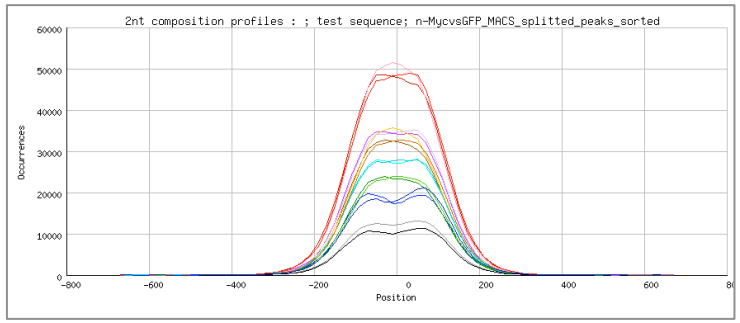
pr1	a	c	g	t	P_P
a	0.29534	0.20296	0.30443	0.19727	0.259
c	0.31618	0.28498	0.07000	0.32884	0.242
g	0.25771	0.24000	0.28381	0.21848	0.24
t	0.17134	0.24103	0.29263	0.29499	0.259



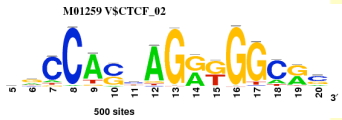
n-Myc



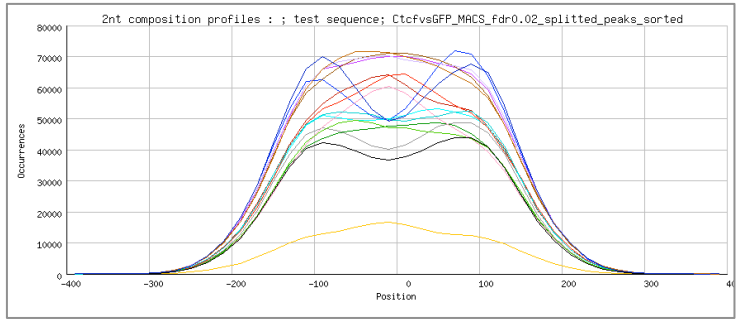
pr1	a	c	g	t	P_I
a	0.22611	0.25100	0.37576	0.14712	0.194
c	0.21417	0.32368	0.22346	0.23869	0.306
g	0.19371	0.32277	0.32413	0.15940	0.306
t	0.12838	0.30586	0.33855	0.22721	0.194

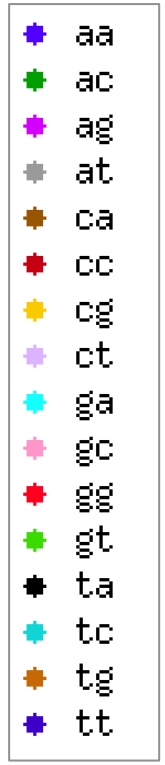


Ctcf



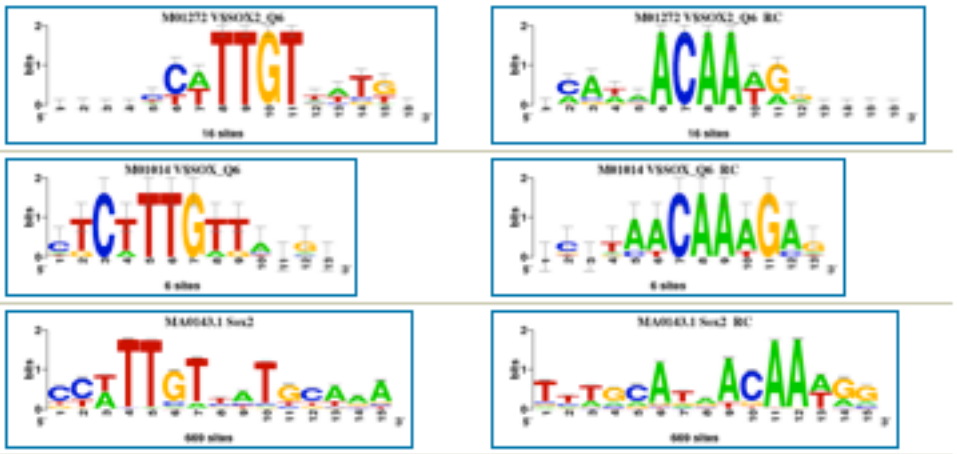
pr1	a	c	g	t	P_I
a	0.28252	0.20757	0.30011	0.20980	0.261
c	0.32450	0.27627	0.06945	0.32979	0.239
g	0.25378	0.24269	0.27546	0.22807	0.238
t	0.18761	0.23215	0.29578	0.28445	0.262





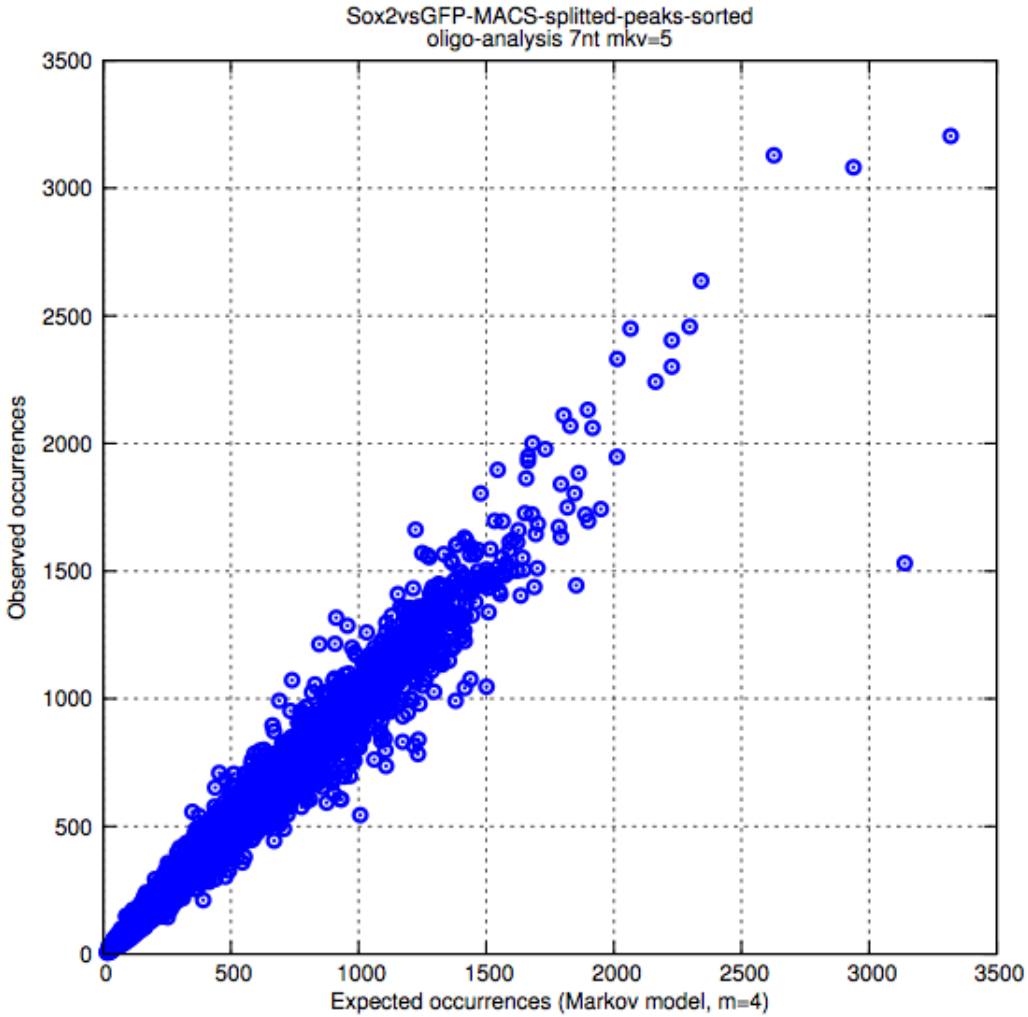
User-specified reference motifs (the "expected" answer)

- 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]

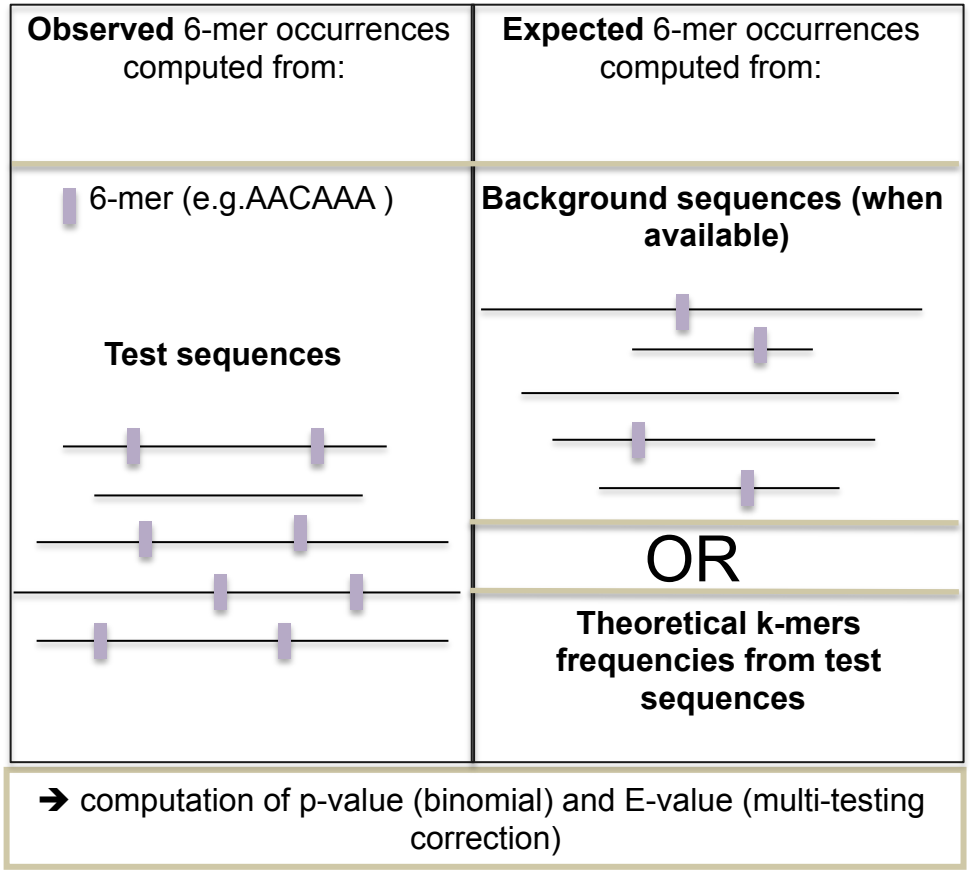
Detection of over-represented oligonucleotides (oligo-analysis)

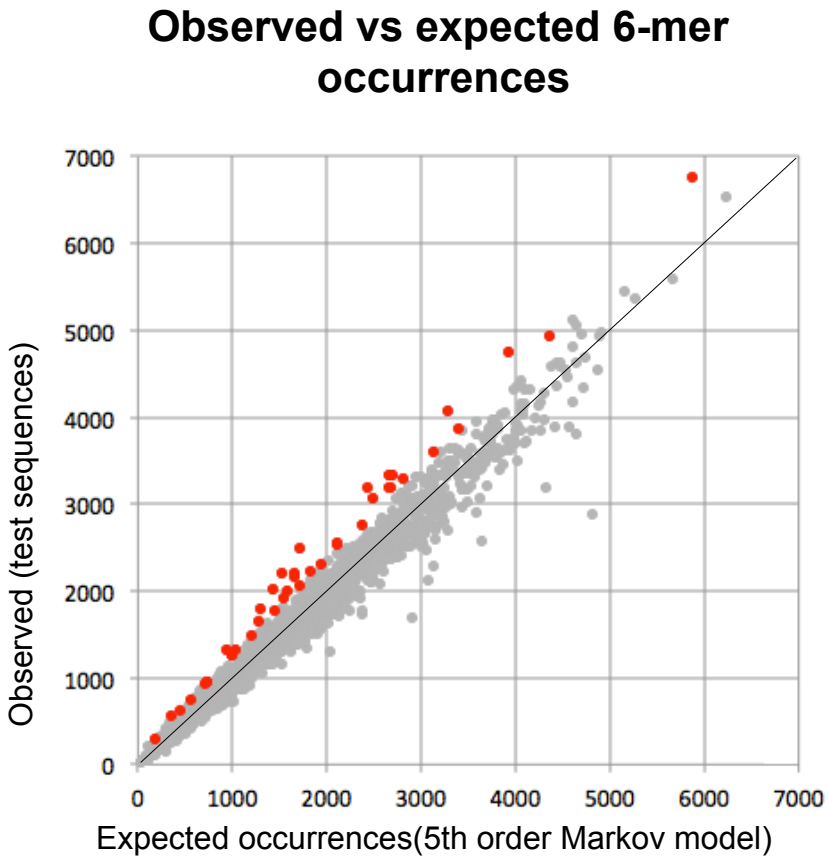
- Principle
  - Count the occurrences of all words (oligonucleotides) of a given size in the input set
  - Estimate the expected number of occurrences according to some background model
  - Report significantly over-represented words.
- Example
  - Sox2 peaks from Chen (2008).
  - Word length *k*=7
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

Detection of global over- or under-representation





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
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|ggggggcg 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
```

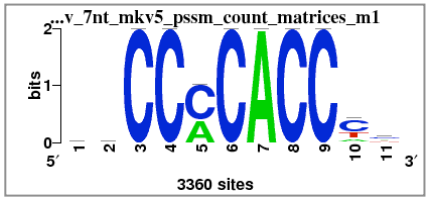
Over-represented words can be assembled to collect matrices

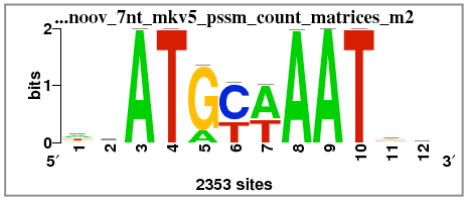
- The list of over-represented words generally contain groups of mutually overlapping words.
- Those groups can be aligned using the program *pattern-assembly*
- Assembled words reveal
  - larger motifs than the initial word length
  - positions with variable residues
- Word assemblies can be used to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

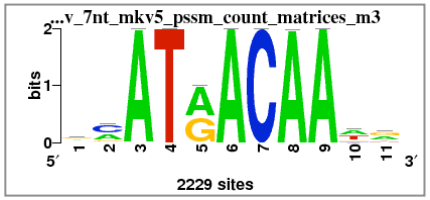
```
;assembly # 1    seed:    2 words length
;alignt rev_cpl score
ccacacc ggtgtgg 31.45
ccccacc ggtgggg 13.34
                31.45    best consensus

;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

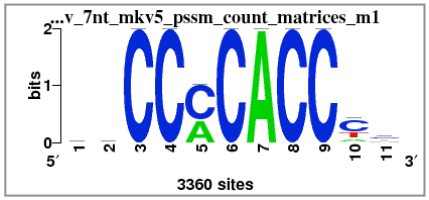
- The significance matrix can be used as “seed” to scan the input sequences and collect binding sites.
- Those sites are in turn used to build a final matrix.

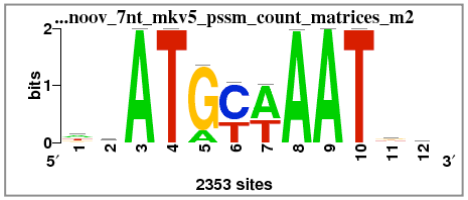
Significance matrix

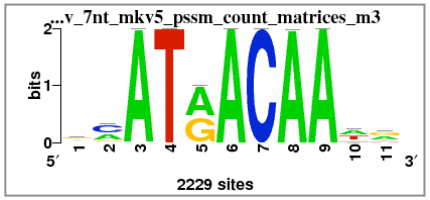
a	0	0	0	0	31.45	0	31.45	0	0	0	0	
c	0	0	31.45	31.45	13.34	31.45	0	31.45	31.45	0	0	
g	0	0	0	0	0	0	0	0	0	0	0	
t	0	0	0	0	0	0	0	0	0	0	0	
//												
a	0	0	27.88	0	19.8	0	27.88	27.88	27.88	0	0	0
c	0	0	0	0	0	27.88	0	0	0	0	0	0
g	0	0	0	0	27.88	0	0	0	0	0	0	0
t	0	0	0	27.88	0	9.36	25.79	0	0	19.8	0	0
//												
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	





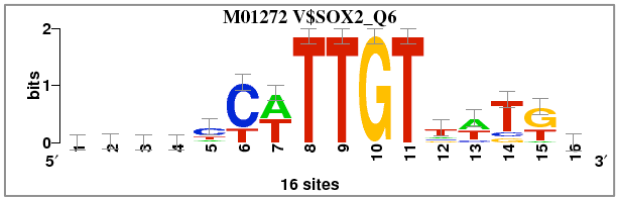


Motifs reported with oligo-analysis (Sox2 peaks from Chen, 2008)

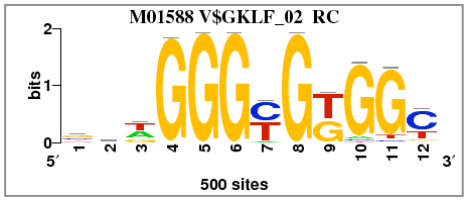
- The program *oligo-analysis* detects over-represented words, as compared to some background model.
- For words of length *k*, we use the most stringent Markov chain model (*m* = *k* – *2*).
- The program detects the Sox2 and Oct4 motifs.
- It also returns a Klf-like motif

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)	

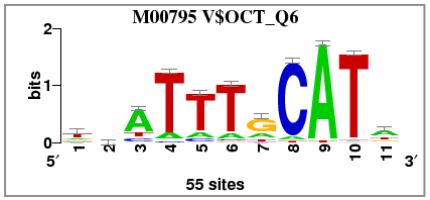
Sox2 (TRANSFAC, built from individual sites)



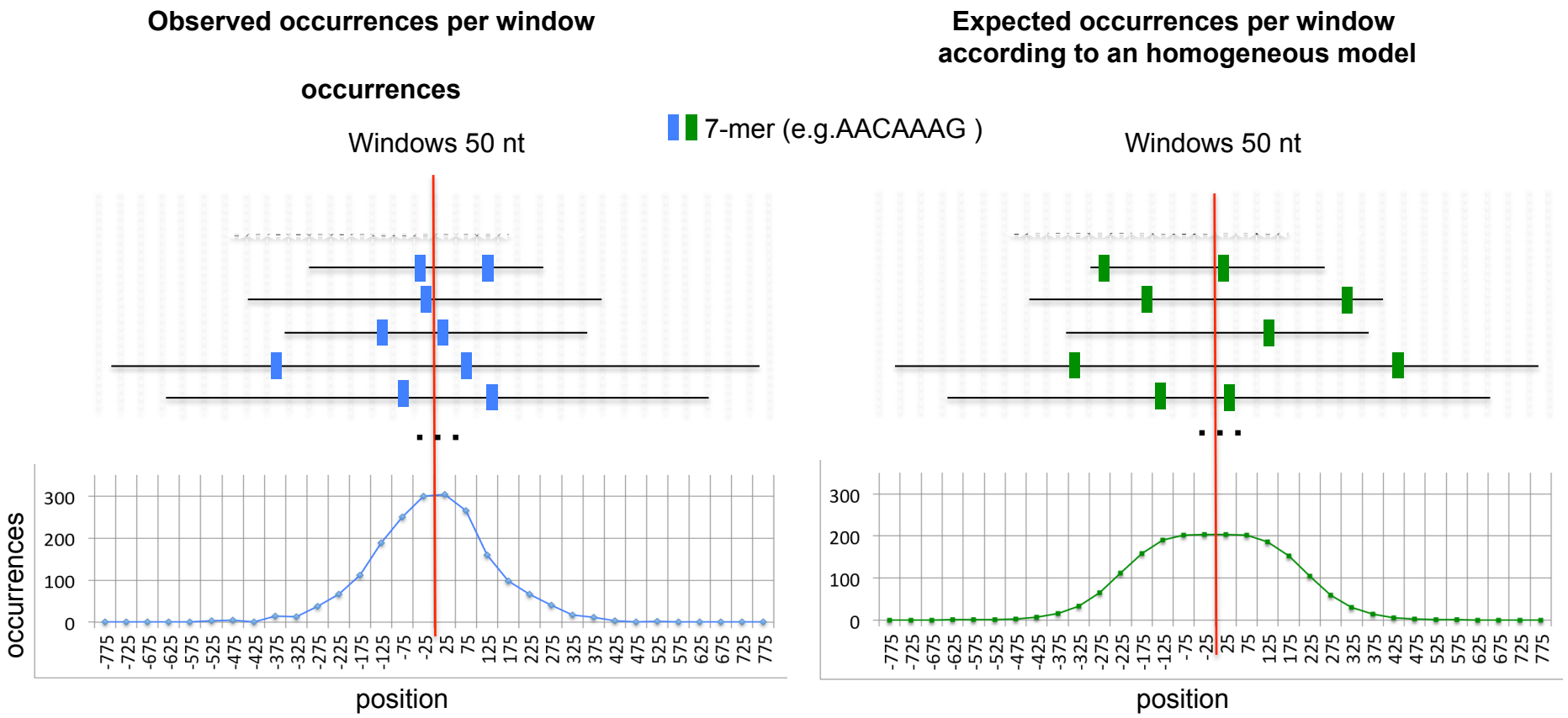
KLF (TRANSFAC built from Chen Klf4 set)



OCT (TRANSFAC, various OCT factors)



Detecting heterogeneous repartition along sequences



Drawing by Elodie Darbo

position-analysis method

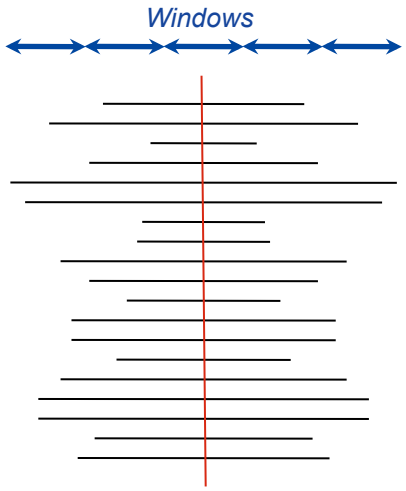
- van Helden, J., del Olmo, M. and Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28, 1000-10.

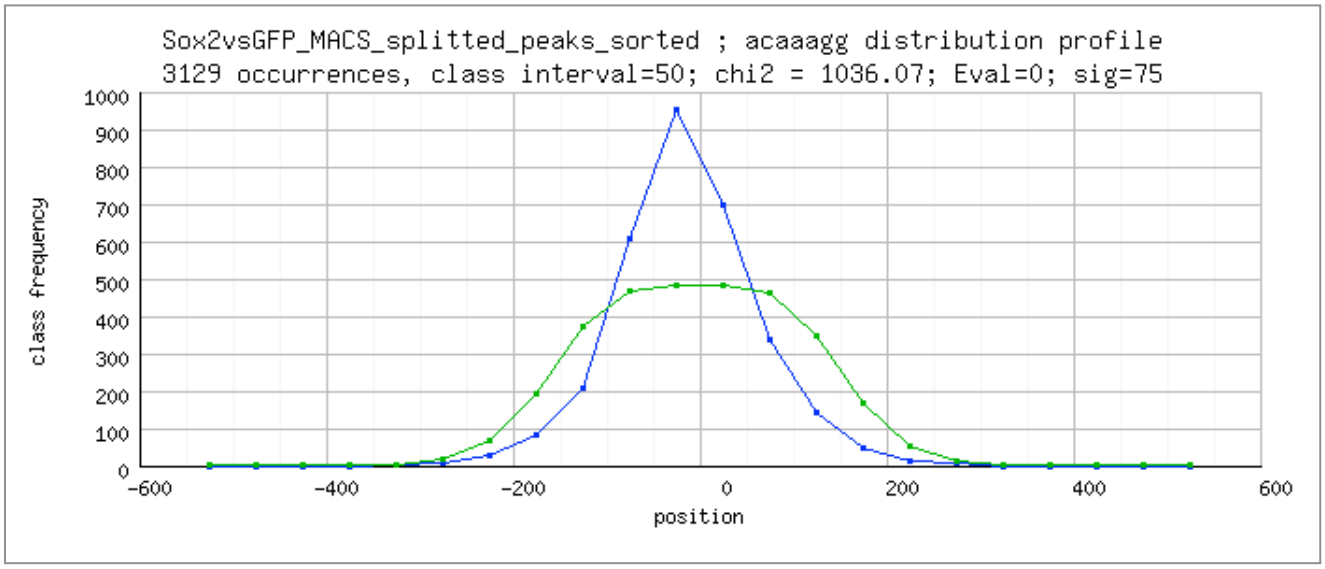
Application to chip-seq:

- 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.
- 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.

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.

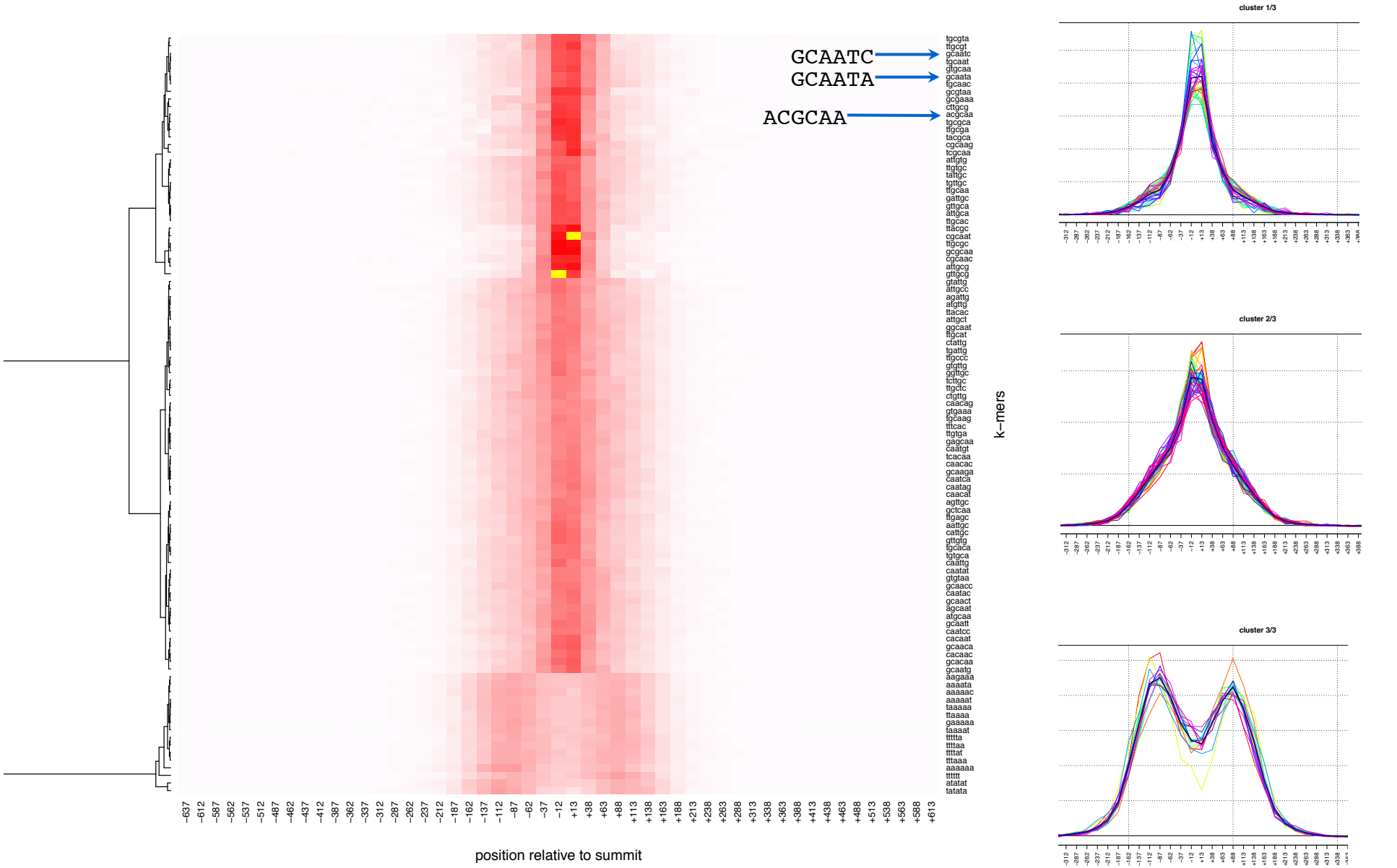




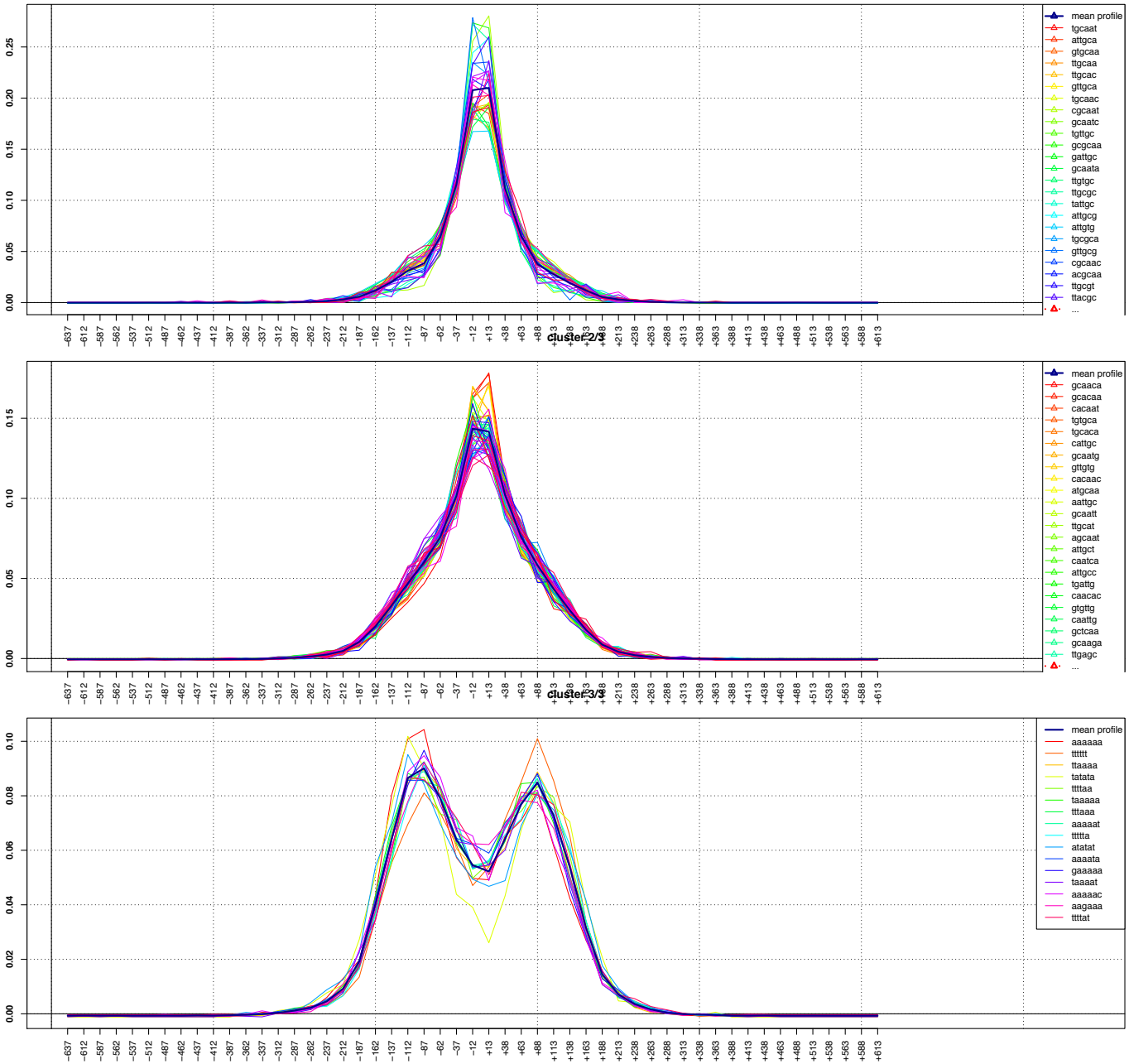
- Green: expected occurrences
  - Note: the expectation decreases with the distance to peak center because peaks have variable lengths.
- Blue: observed occurrences
  - The word ACAAGG is concentrated the center the ChIP-seq peak regions.

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.

Clustering k-mers by positional density profiles



Position profile clustering

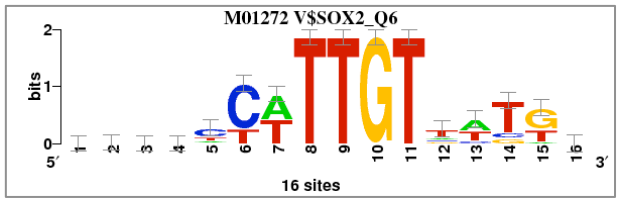


Motifs with position biases in Sox2 peaks from Chen, 2008

- *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)	

Sox2 (TRANSFAC, built from individual sites)

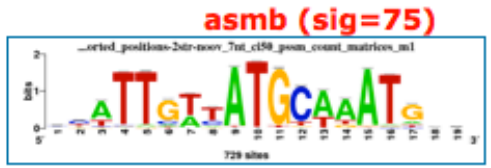


Motifs with position biases in Oct4 peaks from Chen, 2008

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

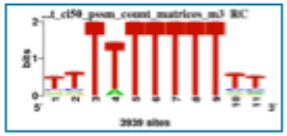
positions_7nt

positions_7nt_m1

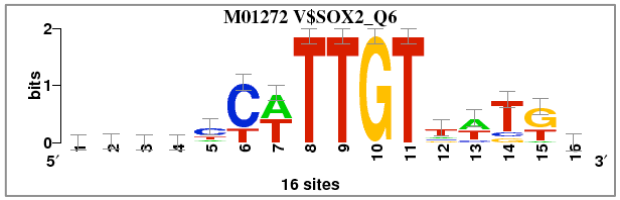


positions_7nt_m2

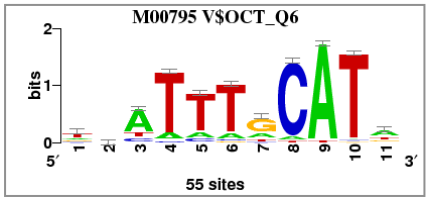
positions_7nt_m3



Sox2 (TRANSFAC, built from individual sites)

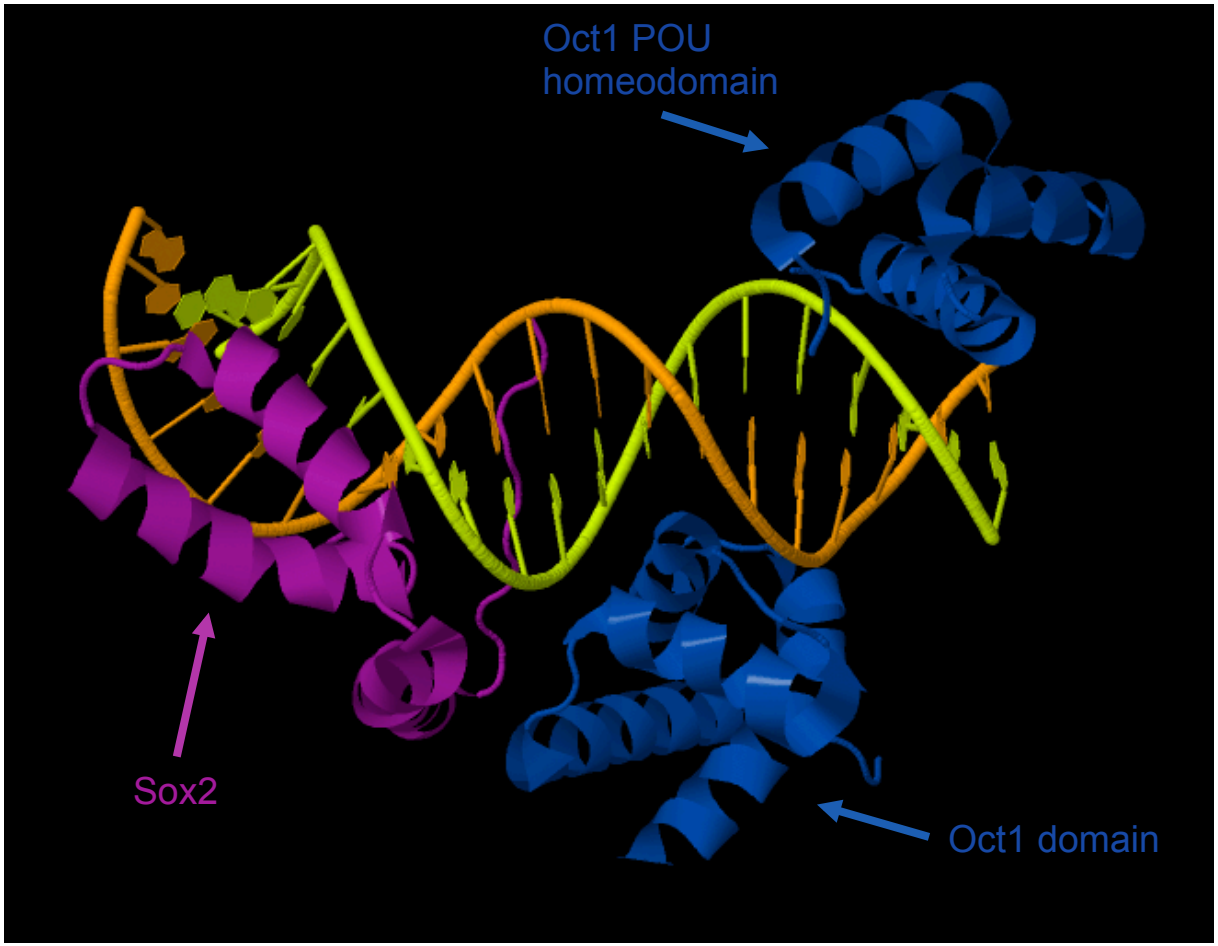


OCT (TRANSFAC, various OCT factors)



Sox2/Oct4 cooperative binding

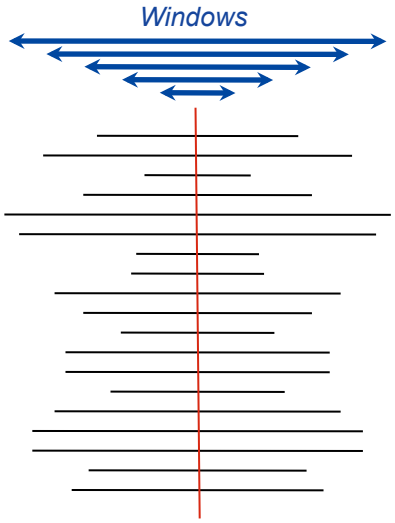
- The Sox2 and Oct 4 transcription factors recognize specific DNA motifs.
- Cooperative binding: Sox2 and Oct4 closely interact to bind DNA.
- The pair of transcription factors recognizes a composite motif called the « SOCT » motif (SOx+OCT).

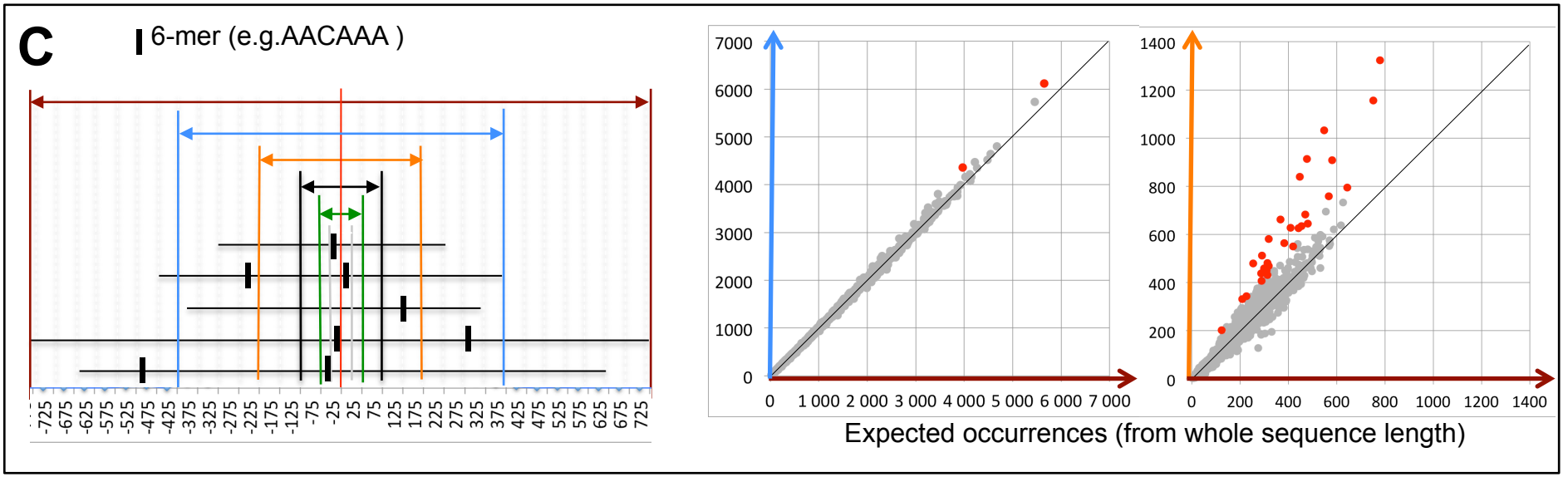


http://www.pdb.org/pdb/explore/explore.do?structureId=1O4X

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.

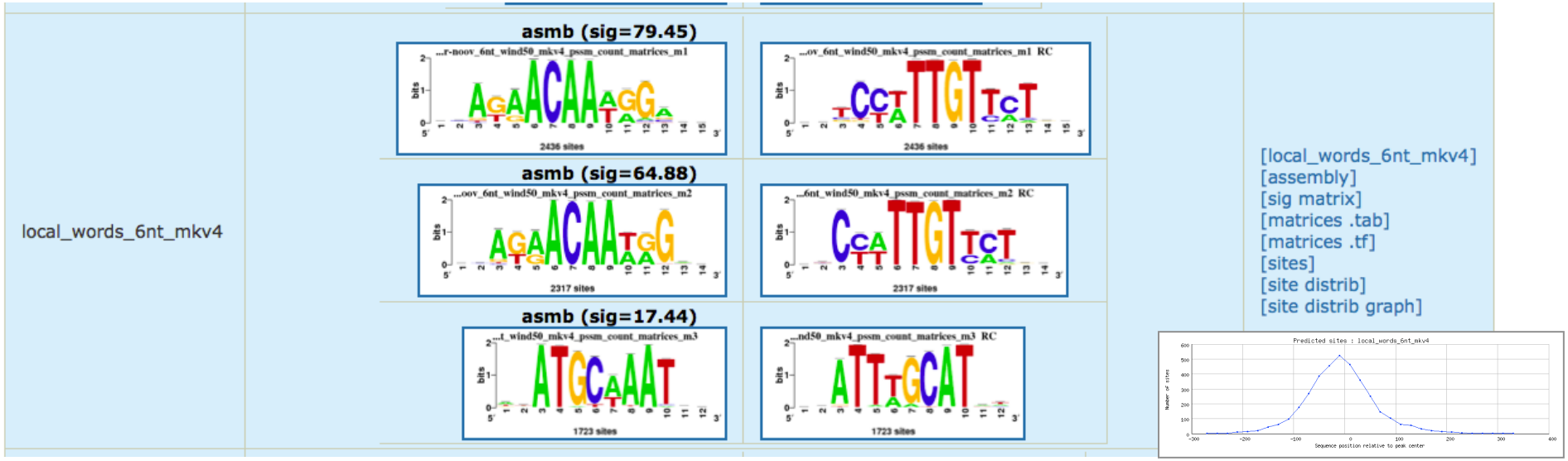




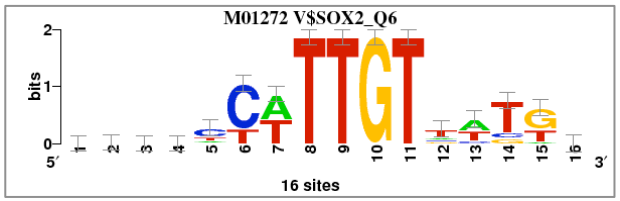
Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. 2012.
A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc 7(8): 1551-1568.

Local over-representation (local-words)

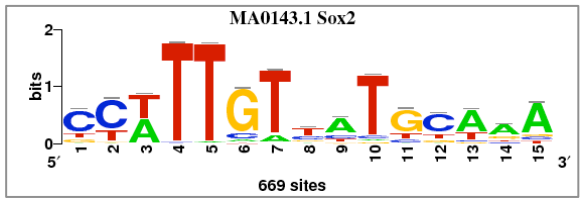
- The program local-words detects windows of local over-representation.
- With windows of 50 bp, the program detects the Sox2 and Oct4 motifs.
- Those motifs are concentrated in the center of the peaks.



Sox2 (TRANSFAC, built from individual sites)

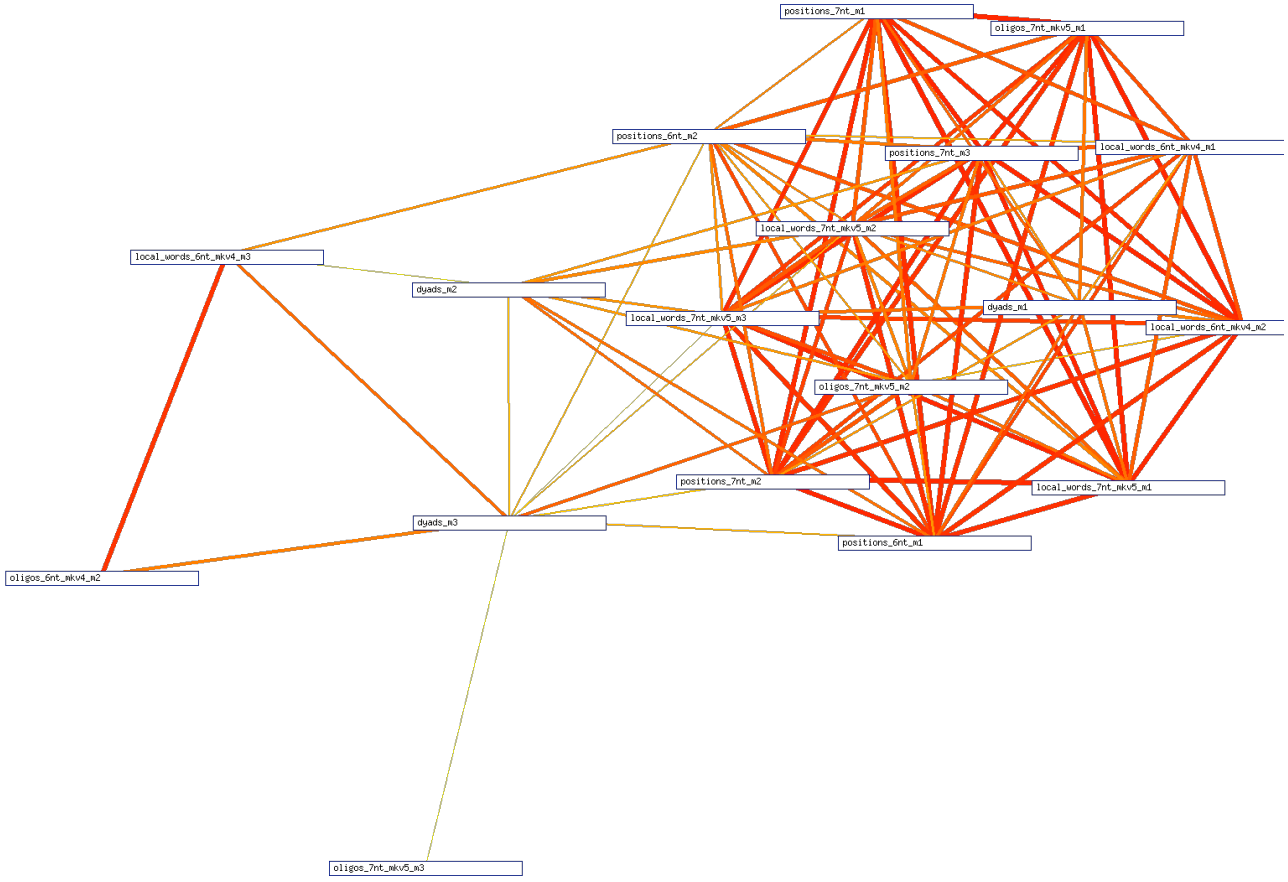


SOCT (JASPAR, built from Chen Sox2 set)



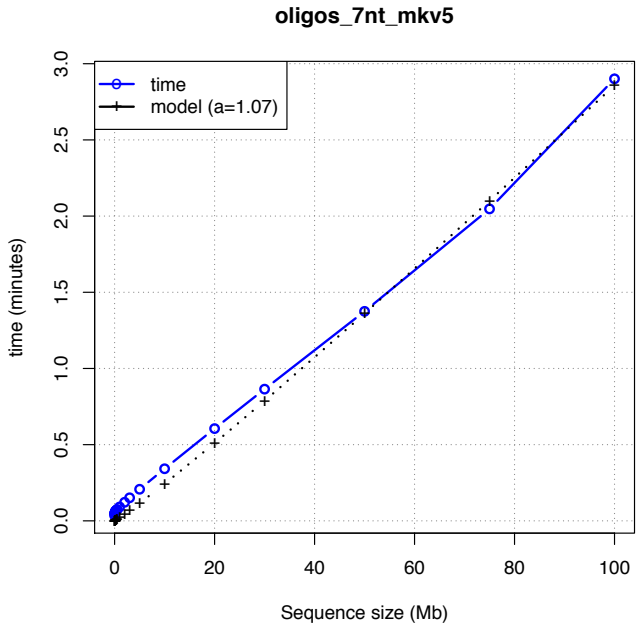
Comparisons between discovered motifs

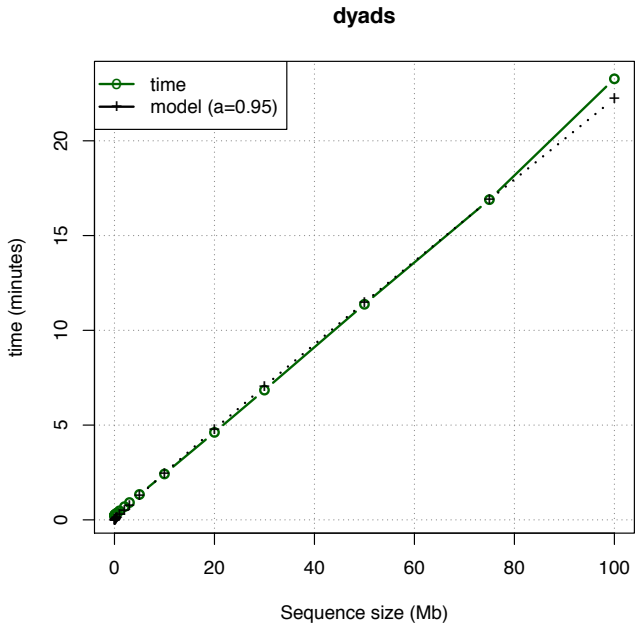
- Pairwise comparisons show the consistency between the motifs discovered by the different approaches.

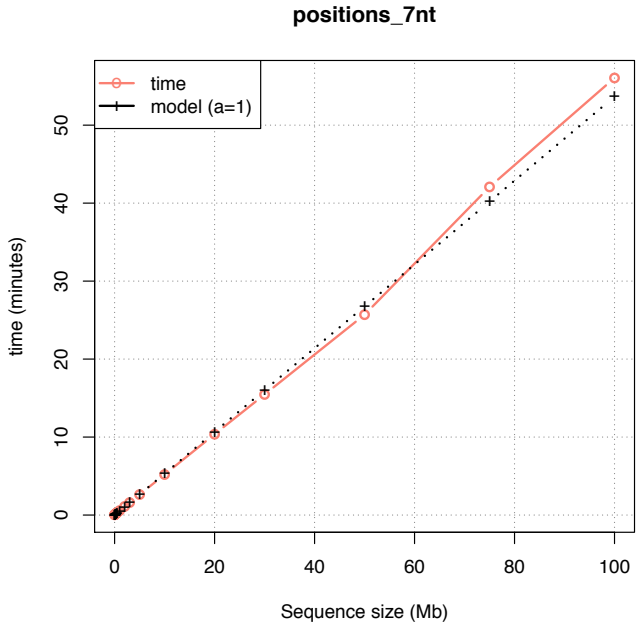


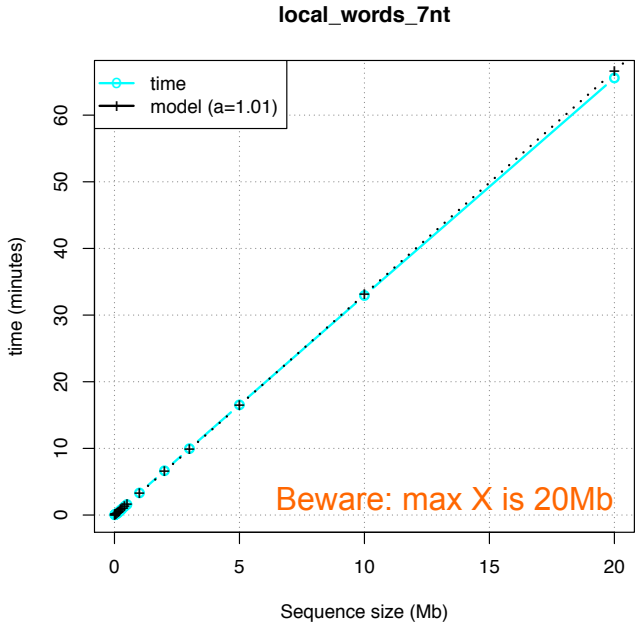
Time efficiency

Time complexity of RSAT word-based algorithms

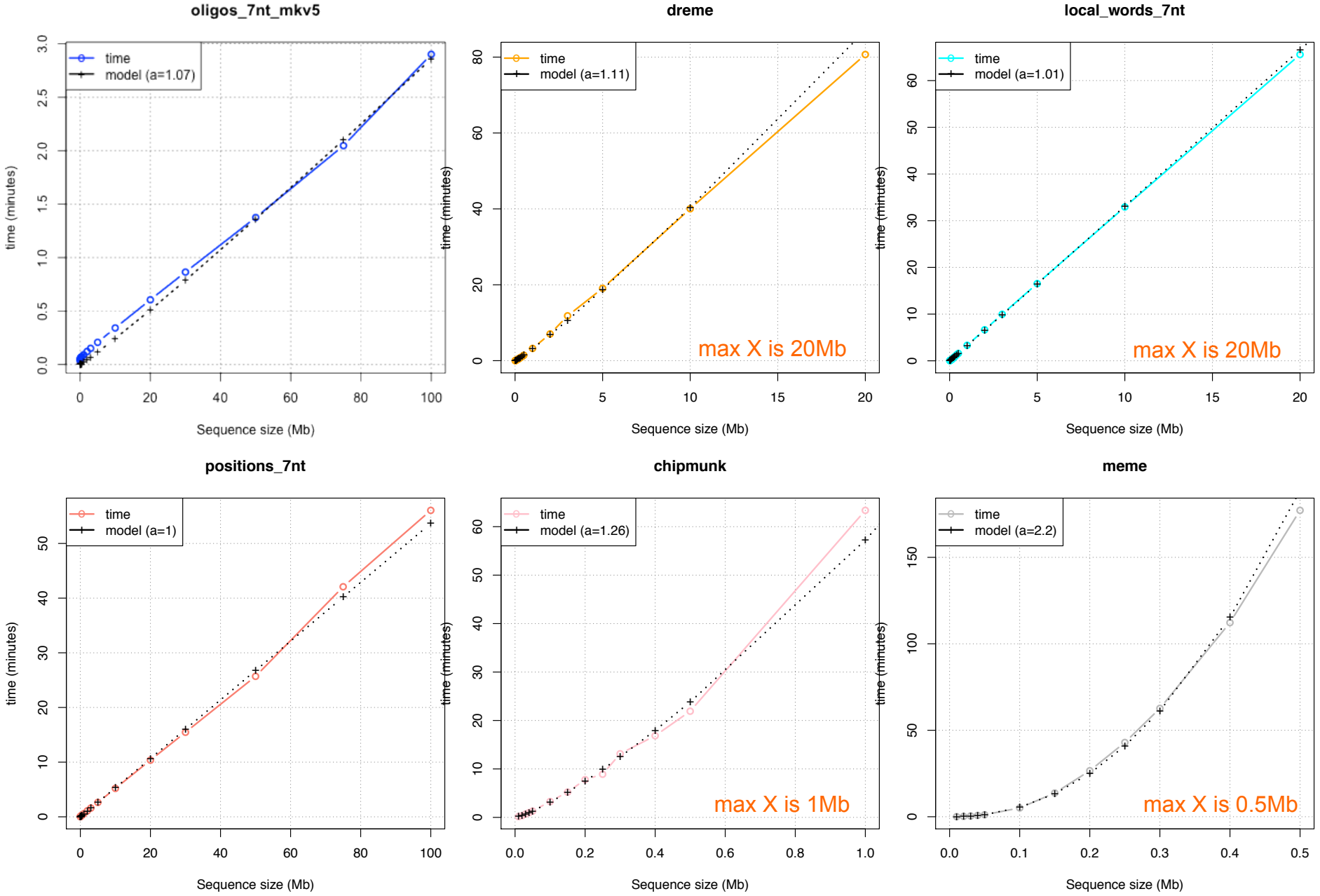




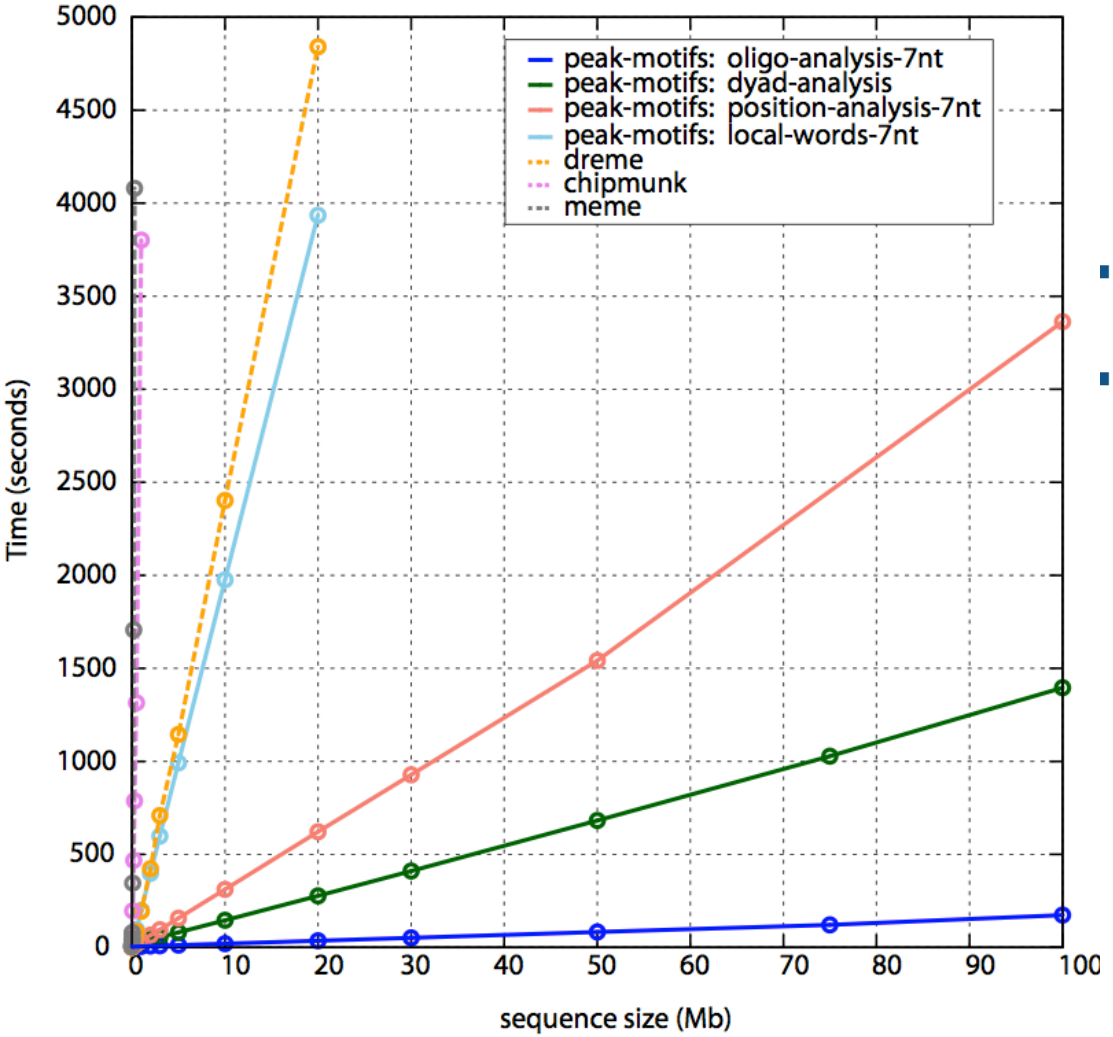




Time complexity



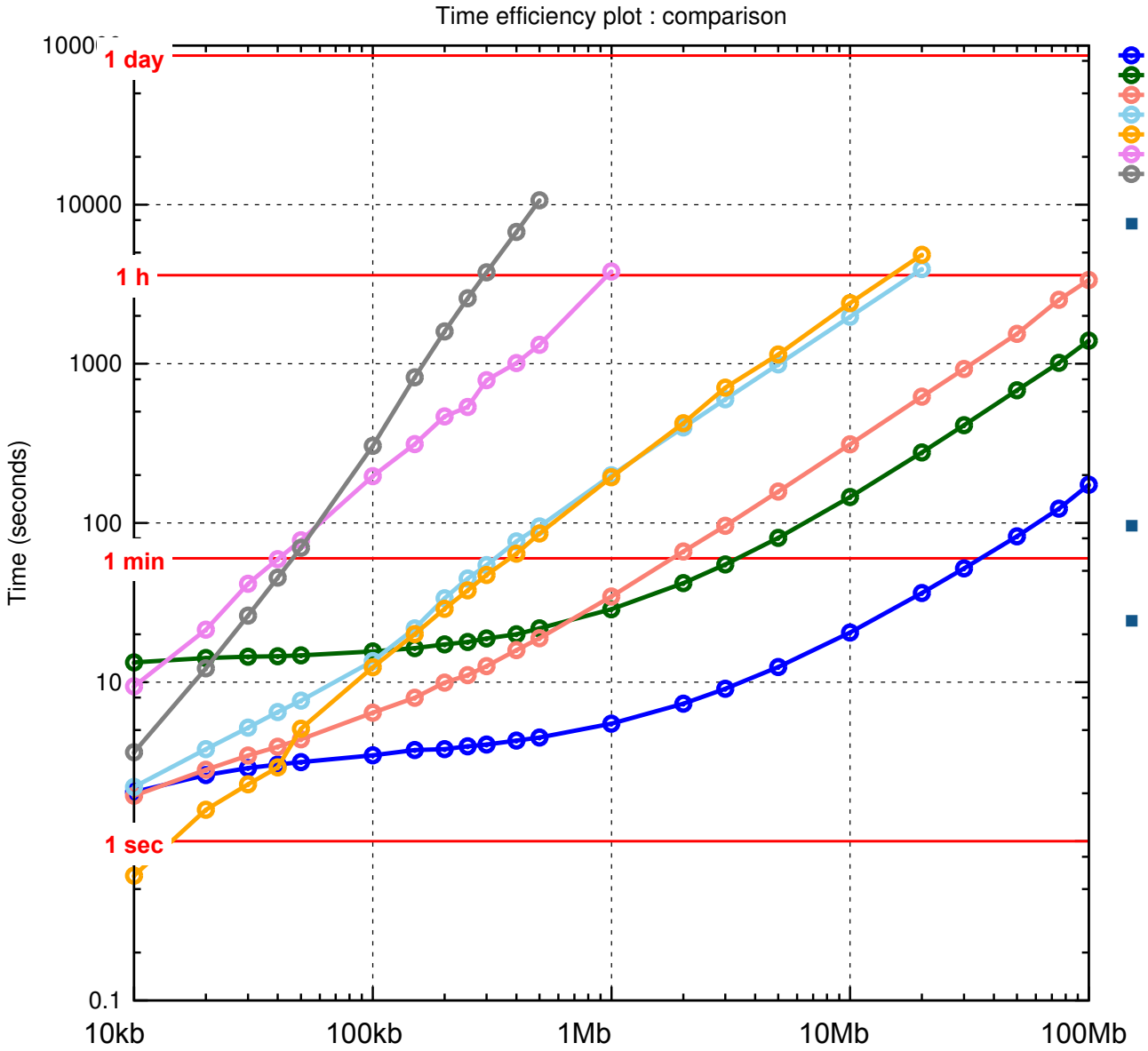
Scalability



- 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 : position-analysis



- 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 complexity (with extrapolations)

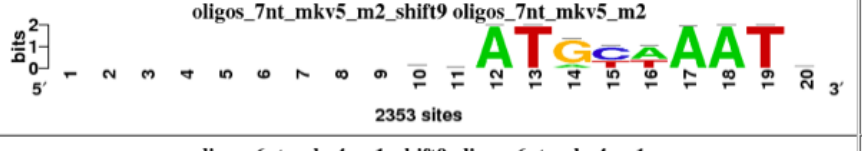
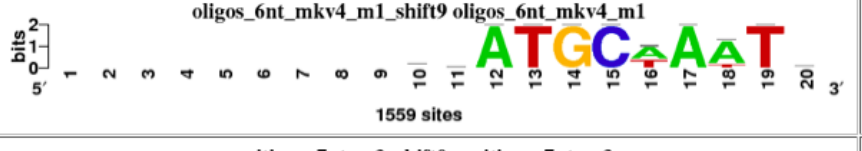
Algorithm	Max tested size (Mb)	Time for max tested size (min)	power (slope of log linear model)	intercept of log linear model	0.1Mb	1Mb	10Mb	100Mb
oligos_7nt_mkv5	100	2.9	1.07	-3.88	0.00	0.02	0.24	2.86
dyads	100	23.27	0.95	-1.29	0.03	0.28	2.47	22.22
positions_7nt	100	56.05	1.00	-0.64	0.05	0.53	5.32	53.65
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*

Relevance of the discovered motifs:
comparisons with
- reference motifs
- database motifs

Discovered versus reference motifs

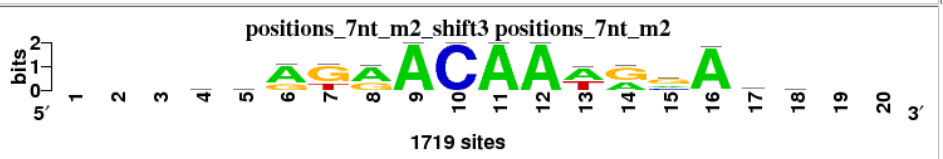
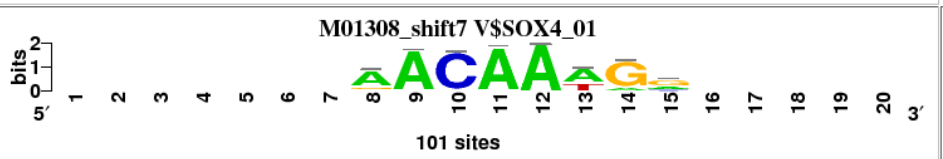
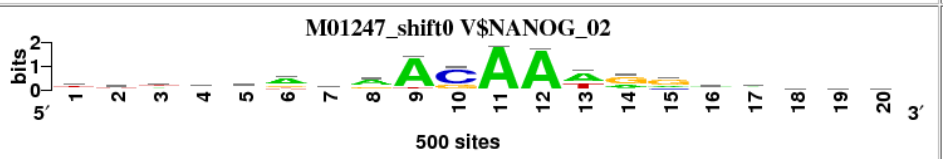
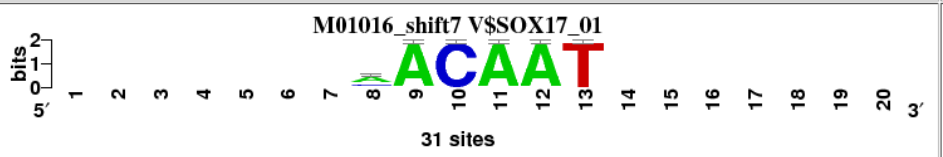
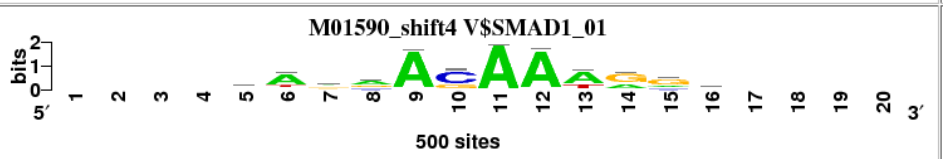
- Discovered motifs are compared to and aligned with the reference motifs.
- The program *compare-motifs*
  - supports various scoring schemes for assessing the similarity between motifs: correlation, Euclidian, Sandelin-Wasserman, SSD, ...
  - Generates multiple (one-to-many) alignment between matrices 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)											
local_words_6nt_mkv4_m3_shift1 (local_words_6nt_mkv4_m3)		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)		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)		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)		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		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).
  - *Note : the database motif for NANOG is wrong: this motif is bound by Sox, not Nanog.*

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

Motifs discovered by peak-motifs in Chen et al. (2008)

Data set	Reference motif	Best-matching discovered motif	Cor	Cov	Returned Reference Motifs	oligos	positions	local-words	dyads	Other motifs found
c-Myc	CACGTGb	rcCACGTGgy	0.99	0.70	V$MYCMAX_03 V$MYCMAX_02 V$CMYC_01	X	X	X	X	LBP1, SP (oligo-analysis) MEF2 (position-anlaysis)
CTCF	ygrCCAsyAGrkGGCrsyr	grCCACyAGrkG	0.95	0.63	V$CTCF_02; V$CTCF_02	X	X	X		MEF2, SP (oligo-analysis) other motifs (dyad-analysis)
E2F1	ttTTTCsCGsc	SSCGGSRGCGSS	0.90	0.50	V$E2F_Q2	X		X		match only covers the right side of the motif
Esrrb	yCAAGGTCAcss	gtTCAAGGTCAkc	0.94	0.79	V$ERR2_01 Jaspar MA0141.1					No results!
Klf4	rCCmCrCCCwkc	rrCCmCrCCCTyy	0.99	0.86	V$GKLF_02	X	X	X	X	E2F, FXR, ER, LBP1 (oligos_6nt)
n-Myc	CACGTG	ryCACGTGry	1.00	0.60	MA0104.1 V$NMYC_01; V$EBOX_Q6_01	X	X	X		AP, SP (oligo-analysis) c-Myc, MEF2 (position-analysis) + other motifs (dyad-analysis)
Nanog	ggGvyCATTkcC	WMATTWSCATTW	0.81	0.50	V$NANOG_01	X				All programs match the Sox2 motifs, only oligo-analysis found the real Nanog motif. Note: the TRANSFAC motif called "V$NANOG_02" is actually a Sox2 motif. It is found by all programs. Additional motifs found with oligo-anlaysis, local-words, and position
Oct4	tdATTTgCATw	HATTWRCATWW	0.96	0.79	V$OCT_Q6; V$OCT4_01	X	X	X	X	The Oct4 motif is clearly found, but some of the discovered instances also match the compositite Sox/Oct motifs.
Smad1										Not a single peak selected by MACS with FDR<=0.2
Sox2	CCwTTGTyaTGcaaA	YWTTGTYATKY	0.93	0.73	**V$SOX2_Q6** V$SOX_Q6	X	X	X	X	Sox but also Oct motifs are found by the different algorithms
Stat3	TTCCaGGAAr	syTTCCwGGAAGtsm	0.98	0.67	V$STAT_Q6; V$STAT_01		X		X	oligo-analysis returns SP, AP, and ER motifs; strangely, dyad-analysis does find the Stat3 motif?!
Tcfcp21	CCrGyyyaadCCrG	hrrArCCAGyyTgr	0.91	0.47	MA0145.1		X			Tcfcp21 found only with position-analysis other motifs returned by other programs
Zfx	ssscAGGCCkcrscss	srGscAGGCCywGss	0.87	0.88	V$ZFX_01	X	X	X	X	AP, SP (oligo-analysis) FOXO1 (positions-analysis) + other motifs (dyad-analysis

- 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.

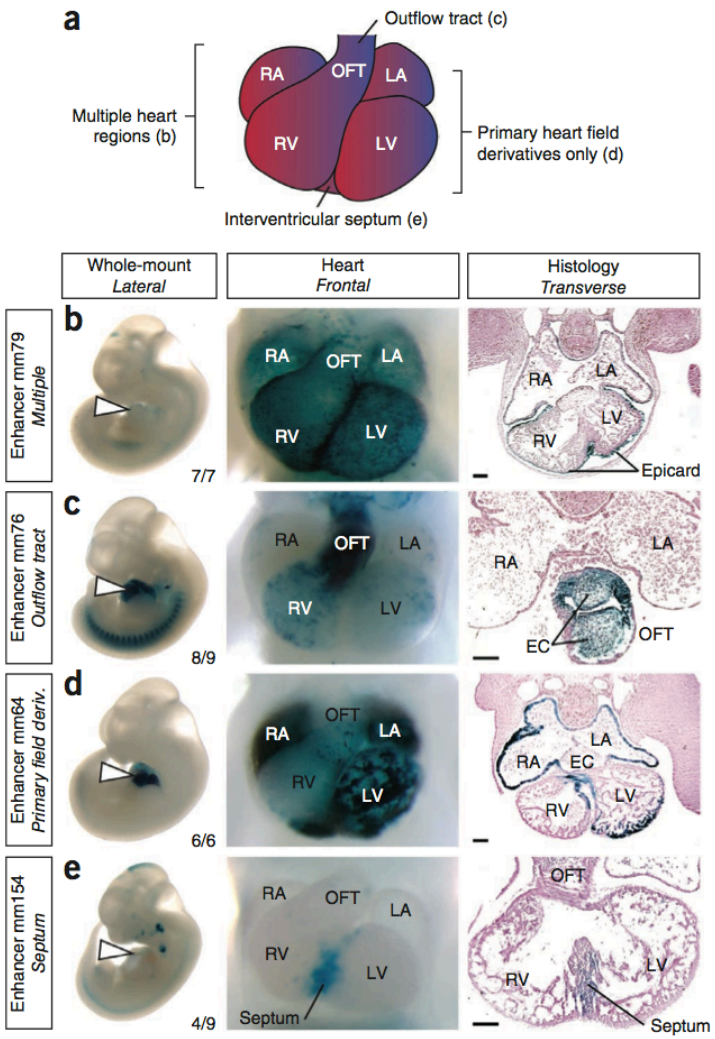
Case study: analysis of p300 binding

Fishing TF binding motifs from transcriptionally active chromomsomal regions

- Blow et al (2010) characterized binding profiles of the histone transacetylase p300 in various tissues during mouse embryonic development.
  - Heart
  - Midbrain
  - Forebrain
  - Limb
- They detected promoter elements involved in the specific activation of gene expression in heart.
- **Bonus:** we could try extracting motifs in the regions bound by p300 in the respective tissues.
  - **Underlying assumption**: alghough p300 is not a motif-recognizing DNA-binding protein, it binds to transcriptionally active regions in the different tissues.
  - **Question:** can we fish out the tissue-specific factors recruiting p300 in those regions ?

ChIP-Seq identification of weakly conserved heart enhancers

Matthew J Blow[1,2], David J McCulley[3,4], Zirong Li[5], Tao Zhang[2], Jennifer A Akiyama[1], Amy Holt[1], Ingrid Plajzer-Frick[1], Malak Shoukry[1], Crystal Wright[2], Feng Chen[2], Veena Afzal[1], James Bristow[2], Bing Ren[5], Brian L Black[3,4], Edward M Rubin[1,2], Axel Visel[1,2] & Len A Pennacchio[1,2]



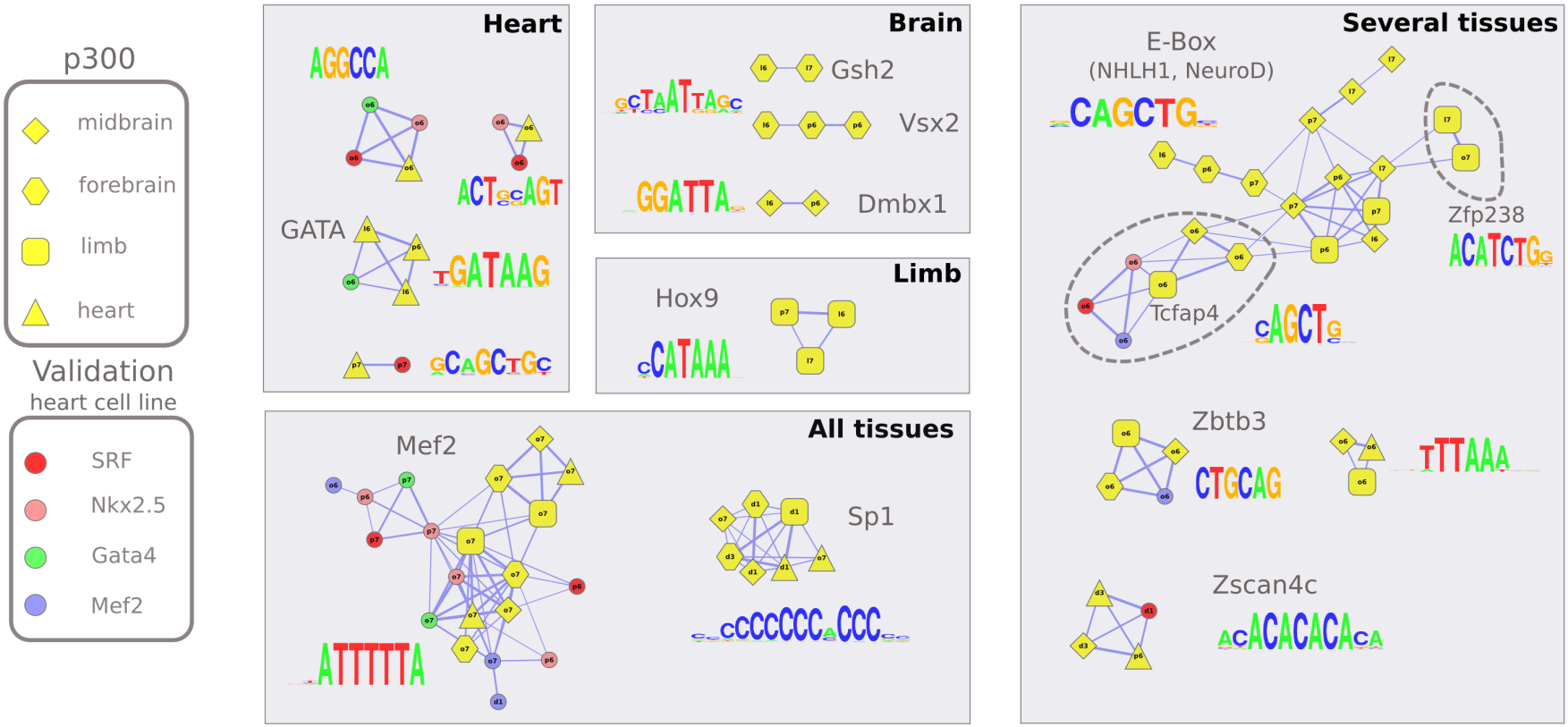
Blow et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet (2010) 42:806-10.

P300 peaks reveal specific TF binding motifs

		Discovered Motif	Algorithm	Best match in motif database	Transcription factor	Expressed in (MGI) …			
						Forebrain	Midbrain	Heart	Limb
Forebrain	1	twkaTTTTTAww	oligos_7nt_mkv5_m1	V$MEF2_05	Mef2C	yes (TS22)	yes (TS22)	**Yes**	**Yes**
	2	aayAAAAACaa	oligos_7nt_mkv5_m2	VSRY_01/VFOXO1/V$CIZ_01	Foxo1,Sry,Zfp384	Yes (brain)	Yes (brain)	?	?
	3	rwCTTTTTAyw	oligos_7nt_mkv5_m3						
	4	ssGGCvsscGsTss	local_words_7nt_m1						
	5	tvTGytaAttaGCAbr	local_words_7nt_m2						
	6	grCAGMTGGys	local_words_7nt_m3						
	7	smgRCAGCTGCygb	positions_7nt_m1	V$NEUROD_02	Neurod1	**Yes**	**Yes**	no	**Yes**
	8	hmtGcTMATKAgCabw	positions_7nt_m2						
	9	wwTTTwAAAww	positions_7nt_m3						
	10	ccCCCCCtCCCCCmc	dyads_m1						
	11	amAAAACAAAam	dyads_m2						
	12	ccCCCCCCCACCCcc	dyads_m3						
Midbrain	1	tyATTTTTAww	oligos_7nt_mkv5_m1	V$MEF2_05	Mef2C	yes (TS22)	yes (TS22)	**Yes**	**Yes**
	2	waCAAAAACaa	oligos_7nt_mkv5_m2	VSRY_01/VFOXO1/V$CIZ_01	Foxo1,Sry,Zfp384	Yes (brain)	Yes (brain)	?	?
	3	cyCCCCCwCCCcc	oligos_7nt_mkv5_m3						
	4	crCCmkCyGCTss	local_words_7nt_m1						
	5	ssaGSmGryGGbg	local_words_7nt_m2						
	6	ssrCAkCTGYss	local_words_7nt_m3	V$NEUROD_02	Neurod1	**Yes**	**Yes**	no	**Yes**
	7	srvCAGCTGbys	positions_7nt_m1	V$NEUROD_02	Neurod1	**Yes**	**Yes**	no	**Yes**
	8	wwwAtATaTwww	positions_7nt_m2						
	9	ssaGCAGMTGGsg	positions_7nt_m3						
	10	cyCCCCCtCCCCCmc	dyads_m1						
	11	amAAAACAAAam	dyads_m2						
	12	caCACACACa	dyads_m3						
Heart	1	arsAAAAACma	oligos_7nt_mkv5_m1	VSRY_01/VFOXO1/V$CIZ_01	Foxo1,Sry,Zfp384	**Yes (brain)**		?	?
	2	ccCCdCCCCCwCCCcc	oligos_7nt_mkv5_m2						
	3	ttATTTTTAaw	oligos_7nt_mkv5_m3	V$MEF2_05	Mef2C	Yes (brain)	Yes (brain)	**Yes**	**Yes**
	4	waTGTTAACAtw	positions_7nt_m1	MA0031.1/V$HNF3A_01	forkhead family			yes (Foxa1;TS24/Foxa2;TS16)	no
	5	cgCsCGCGsGcg	positions_7nt_m2						
	6	CGCSGCGCGCG	positions_7nt_m3						
	7	ccCCCCCCCmCCCcs	dyads_m1						
	8	awAAAAATAAAwa	dyads_m2						
	9	ACACACACACA	dyads_m3						
Limb	1	ttrTTTTTAww	oligos_7nt_mkv5_m1	V$MEF2_05	Mef2C	yes (TS22)	yes (TS22)	**Yes**	**Yes**
	2	awyAAAAACaa	oligos_7nt_mkv5_m2	VSRY_01/VFOXO1/V$CIZ_01	Foxo1,Sry,Zfp384	Yes (brain)	Yes (brain)	?	?
	3	ssCGCCCCCGcs	oligos_7nt_mkv5_m3						
	4	rrcCATAAAAhh	local_words_7nt_m1	V$HOXD13_01	Hoxd13	no	no	?	**Yes**
	5	csGCrGCyGCsg	local_words_7nt_m2						
	6	rmACATCTGkw	local_words_7nt_m3	V$RP58_01	Zfp238	yes (TS17&21)	**Yes**	no	**Yes**
	7	sctsCAGCTGsags	positions_7nt_m1	V$NEUROD_02	Neurod1	**Yes**	**Yes**	no	**Yes**
	8	wwTATATATAww	positions_7nt_m2						
	9	rwcCATAAAAhw	positions_7nt_m3						
	10	ycTCCCCCTCCCCChc	dyads_m1	V$ZFP281_01	Zfp281	no	no	yes (TS17)	**Yes**
	11	wwTAAAATAAAAChm	dyads_m2	V$HOXA13_02	Hoxa13	yes (TS17)	no	yes (TS28)	**Yes**
	12	hmATAAATaw	dyads_m3						

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.

Network of motifs discovered in tissue-specific p300 binding regions



Carl Herrmann
(TAGC, Marseille, France)

ChIP-seq analysis (peak-motifs, compare-matrices).

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.

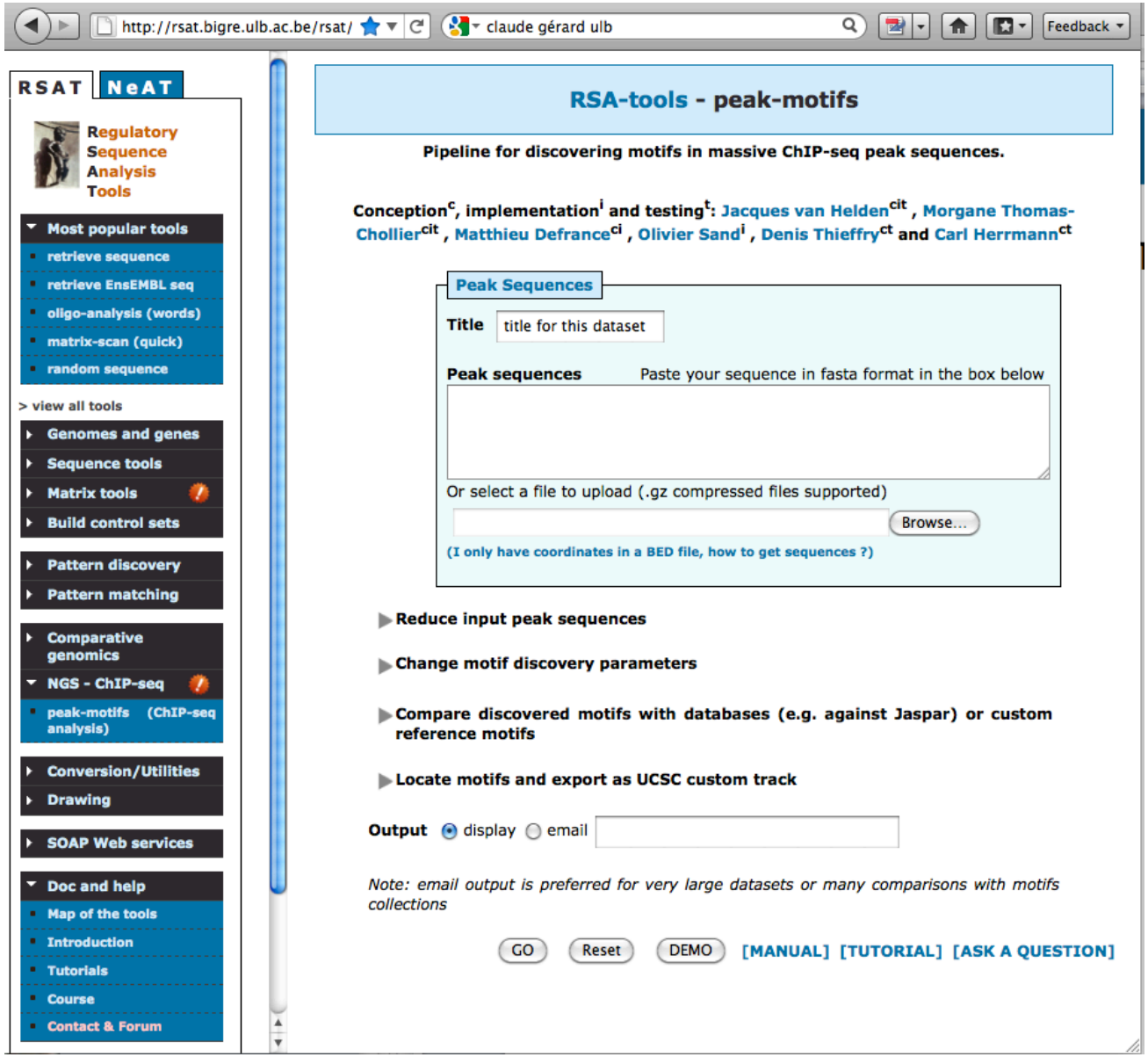
Conclusions

Conclusions

- Results
  - ChIP-seq experiments with specific TFs
    - The correct motif is generally recovered consistently by all algorithms
    - Binding motifs of co-factors are also detected.
  - ChIP-seq experiments with generic factors (e.g. p300)
    - Peak-motifs discovers binding motifs for specific TFs expressed in the tissues.
- The program ***peak-motifs*** provides a flexible tool for analyzing motifs in large collections of peaks.
  - Time-linear algorithms.
  - Low 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
  - Loading of peaks and predicted sites in the UCSC genome browser
- Challenges
  - Integrating genome-wide location profiles from multiple experiments
  - Integrating TF binding and chromatin accessibility profiles
  - Building multi-layer networks by integrating protein/DNA binding with transcription profiles, protein interactions, signaling, metabolism
  - ...
  - Bridging the gap from large-scale, probably incomplete and noisy networks to dynamical models.

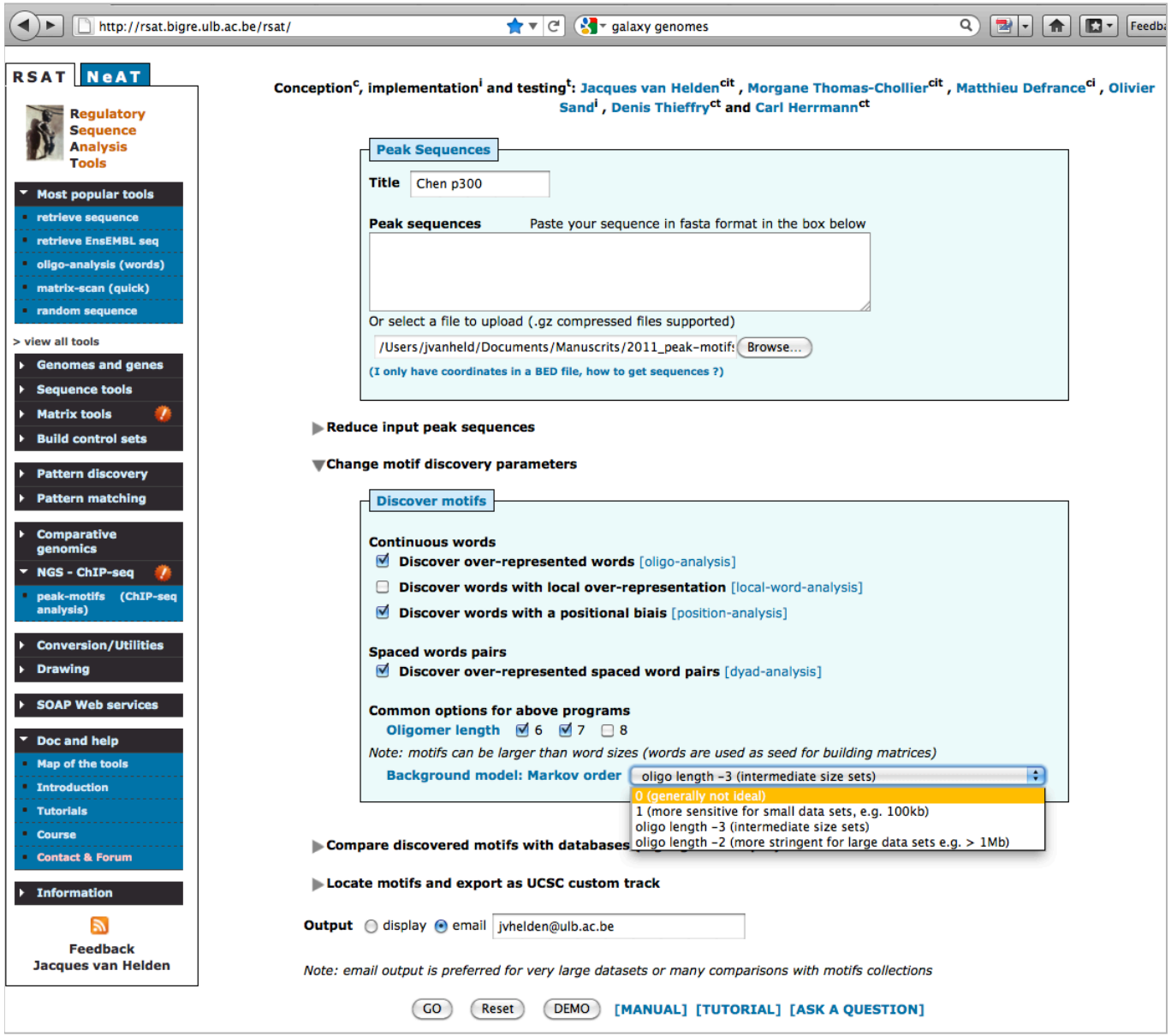
Availability

Availability



- Regulatory Sequence Analysis Tools (RSAT)
  - http://rsat.ulb.ac.be/rsat/
- Interfaces
  - Stand-alone apps
  - Web site
  - Web services (SOAP/WSDL API)
- Web interface
  - Simplcity of use (“one click” interface).
  - Advanced options can be accessed optionally.
  - Allows to analyze data set of realistic size (uploaded files).

Tunability



Regulatory Sequence Analysis Tools (RSAT)

- http://rsat.ulb.ac.be/rsat/

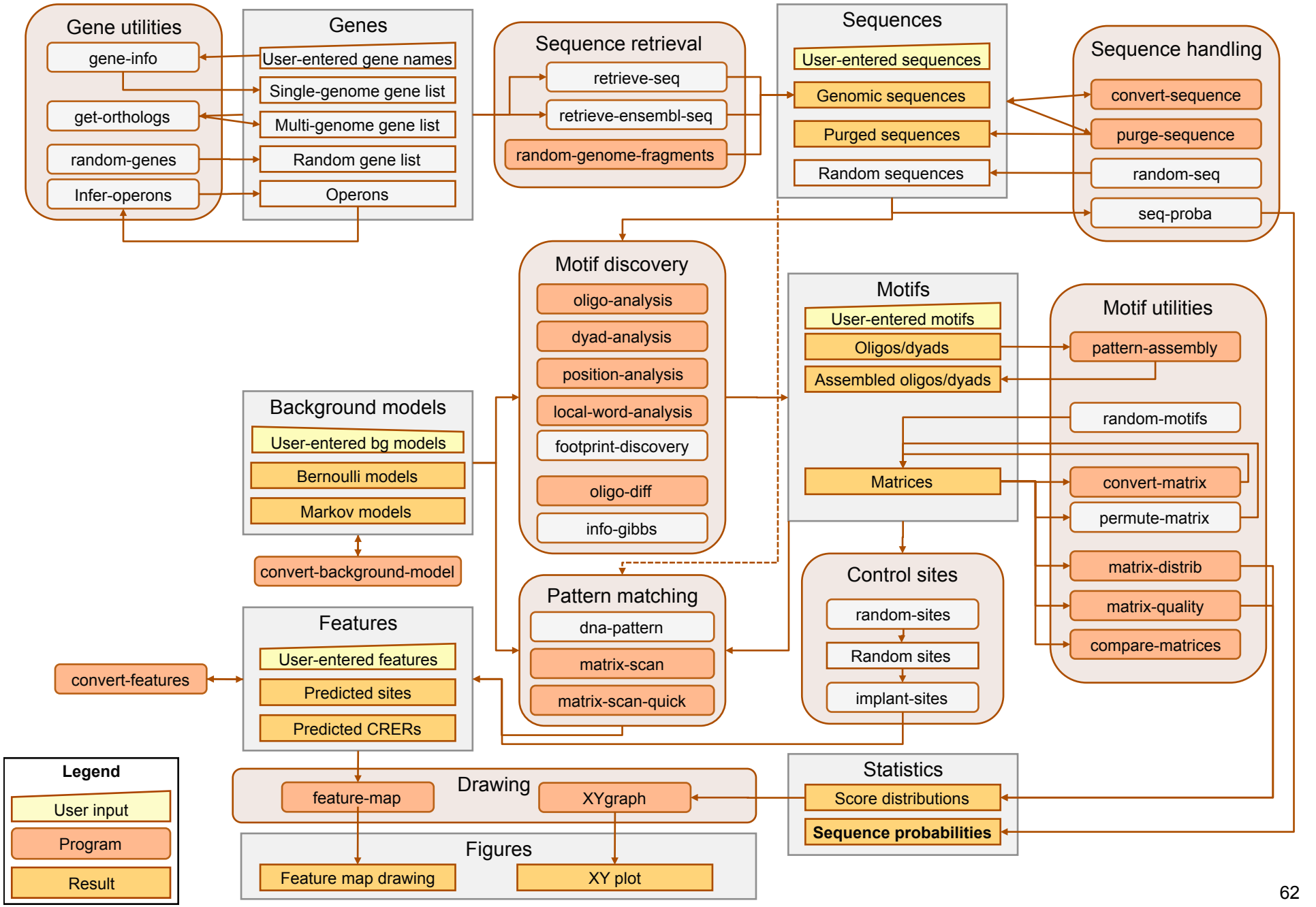
Web interface

- 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



People involved in the approaches and results presented today

Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Regulatory Sequence Analysis Tools development team

(RSAT, http://rsat.ulb.ac.be/rsat/)

- **Morgane Thomas-Chollier (Max Planck, Berlin)**
- **Olivier Sand (Lille, France)**
- **Matthieu Defrance (Faculty of Medicine, ULB)**
- **Alejandra Medina-Rivera (UNAM – Mexico)**

Main collaborators

- **Denis Thieffry (TAGC, France)**
- **Carl Herrmann (TAGC, France)**
- Elodie Darbo (TAGC, France)
- Julio Collado-Vides (UNAM, Mexico)
- Bruno André (ULB, Belgium)
- Fadi Abdel-Sater (ULB, Belgium)
- Cei Abreu-Goodger (Sanger, UK)

Former project participants

- Rekin's Janky
- Jean-Valéry Turatsinze
- Eric Vervisch
- Nicolas Simonis

Mobile Genetic Elements

(ACLAME, http://aclame.ulb.ac.be/

- Raphaël Ieplae
- Gipsi Lima-Mendez
- Ariane Toussaint

Network Analysis Tools

(NeAT, http://rsat.ulb.ac.be/neat/)

- Sylvain Brohée

Metabolic path finding

(NeAT, http://rsat.ulb.ac.be/neat/)

- Karoline Faust
- Didier Croes

Former project participants

- Shoshana Wodak
- Fabian Couche

Programs from external developers

- Pierre Dupont (UCL)

Acute Lymphoblasitc Leukemia networks

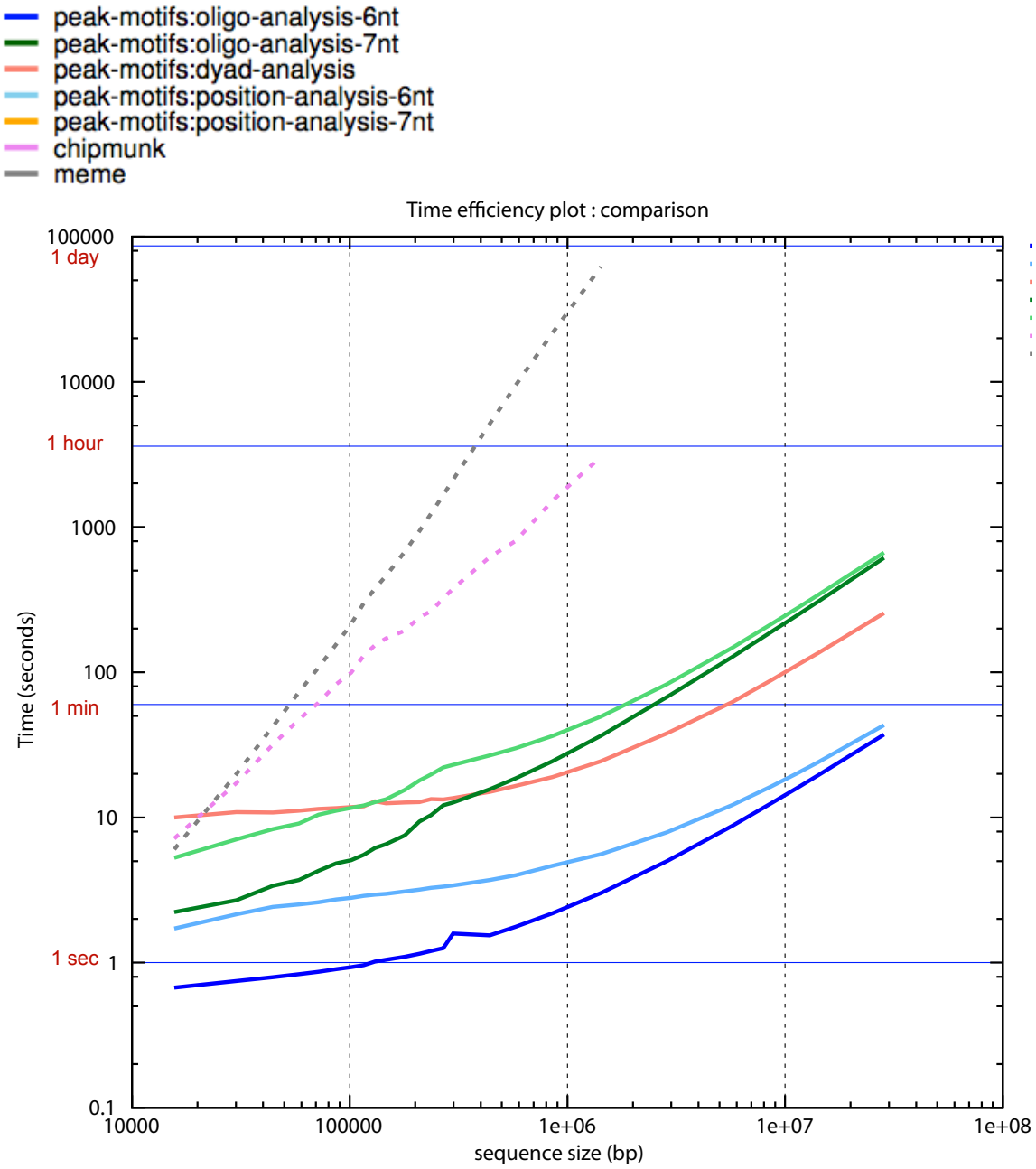
- Nicolas Simonis
- Léon Juvénal Hagingambo

Collaborators

- Jean-Claude Twizere (Fac. Gembloux)

Supplementary material (not shown)

Time efficiency : times for treating Chen data sets



- MEME
  - Processing time is quadratic.
    - 10 hours for 1Mb
    - 1000 hours for 10Mb
- 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.
  - 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.

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