Ecole de Bioinformatique Aviesan (EBA), Roscoff, Sept-Oct 2015

Discovering motifs in ChIP-seq peaks

with the Regulatory Sequence Analysis Tools (RSAT, http://rsat.eu/)

Jacques van Helden, Morgane Thomas-Chollier,
Matthieu Defrance, Carl Herrmann, Stéphanie Legras

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)
http://www.bigre.ulb.ac.be/

Previous sessions

- Read mapping: from raw reads to aligned reads.
- Peak calling: from aligned reads to regions/peaks of high read density.
- ChIP-seq annotation
  - Identification of genes related to the peaks.
  - Profiles of ChIP-seq reads around reference points (TSS, histone marks, …).
  - Functional enrichment of the genes related to the peaks.

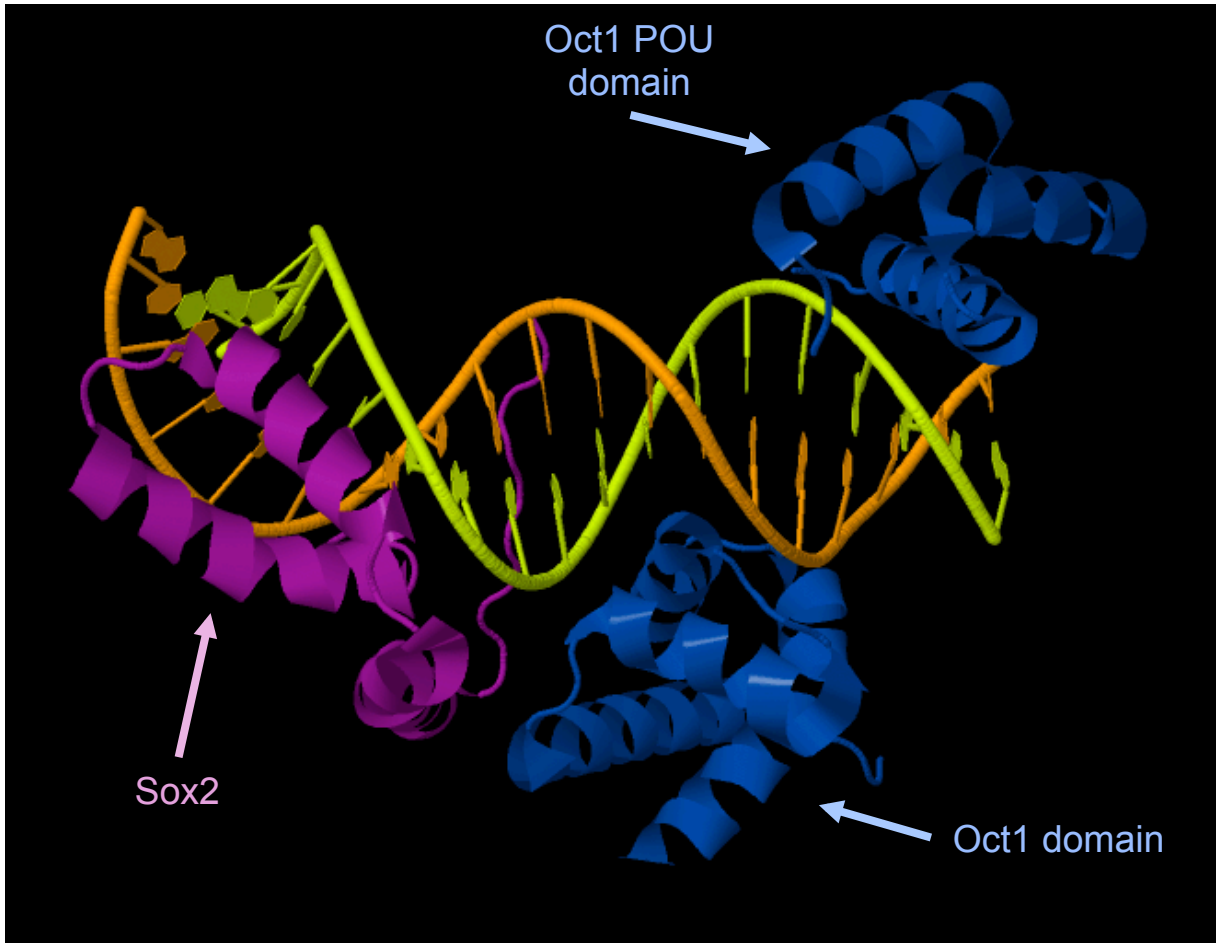
Motif analysis in ChIP-seq peaks: what for ?

- ***Motif discovery*** from peak sequences, without a priori (*"de novo" analysis*).
  - Check if the ***expected motif*** (ChIP-ped factor) can be discovered from the peaks.
    - If not, evaluate if the experiment and bioinformatics treatment was OK (e.g. functional enrichment).
  - ***Improve annotated motifs***
    - Obtain a well-documented motifs (built from thousands of sites), supposedly more reliable than "classical" motifs build from individual experiments (e.g. 10 sites from footprints and EMSA).
    - Main annotation path for recent motif database releases (JASPAR, TRANSFAC, …).
  - Discover ***partner transcription factors***.
- ***Differential motif discovery***
  - Discover differentially represented motifs between a peak set of interest (*test*) compared to another one (*control*).
- ***Peak scanning***
  - Goal: identify binding sites within the peaks
  - Typical ChIP-seq peak: ~100 to 1000bp Actual binding site: 6 to 10 bp
- ***Peak enrichment*** for known motifs
  - Scan sequences to identify putative binding sites for TFs known to interact.
  - Compare observed/expected number of sites

From transcription factor binding sites (TFBS) to motifs (TFBM)

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).

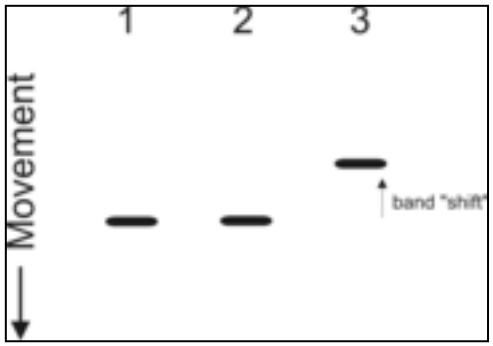


http://www.pdb.org/pdb/explore/explore.do?structureId=1O4X

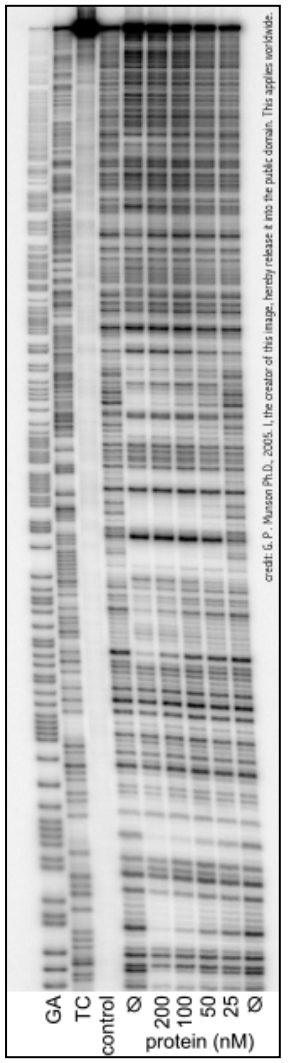
Transcription factor binding site prediction : difficulties

- 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.

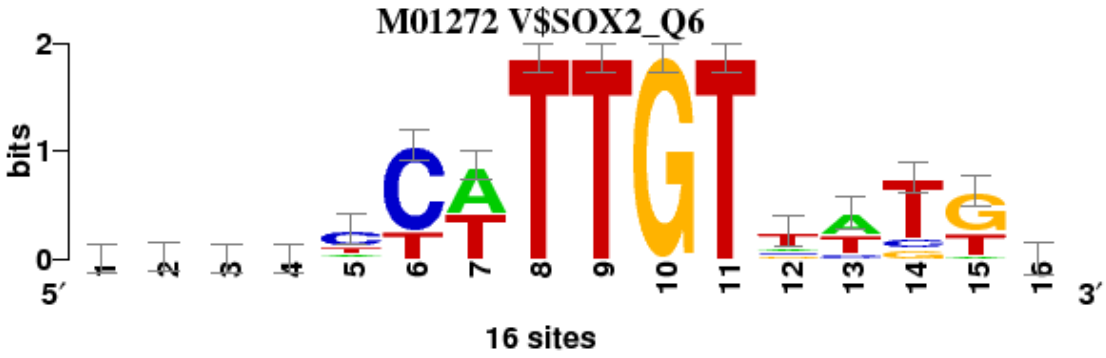
Collection of binding sites
used to build the Sox2 matrix
(TRANSFAC M01272)

```
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, ...).

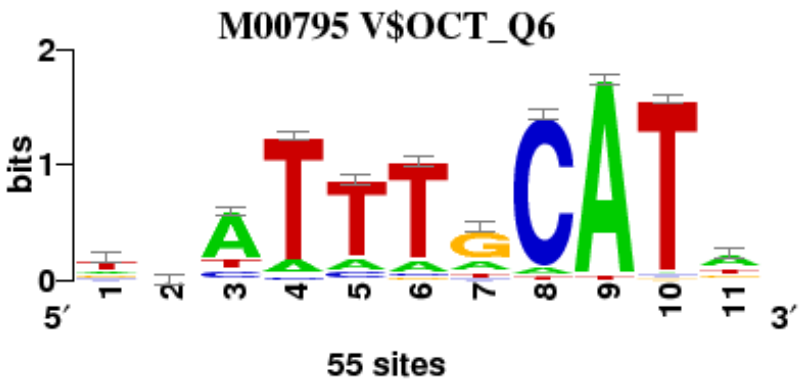
Collection of binding sites
used to build the motif of the OCT ***family*** *(TRANSFAC M00795)*

```
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



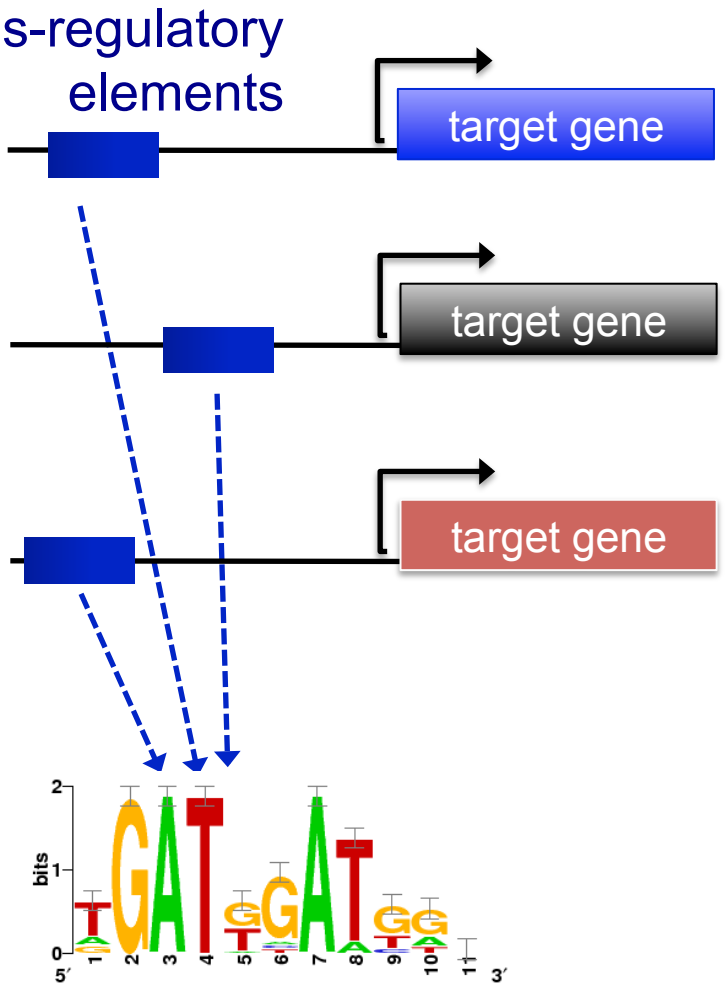
Motif discovery

"De novo" analysis: from sequences to motifs

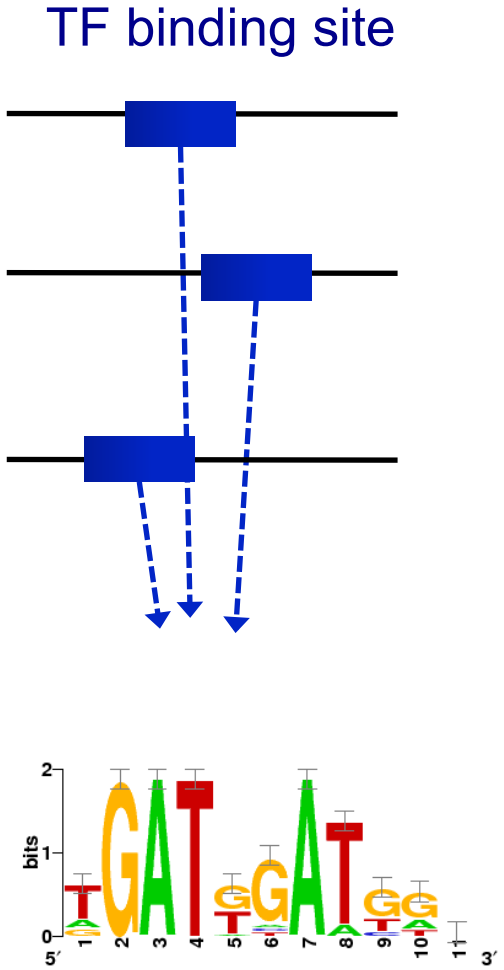
Motif discovery (de novo)

- Assuming that a transcription factor binds most of the genomic regions obtained from an experiment, can we discover its motif on the basis of these sequences only?

Case 1: promoters of co-expressed genes



Case 2: ChIP-seq peaks

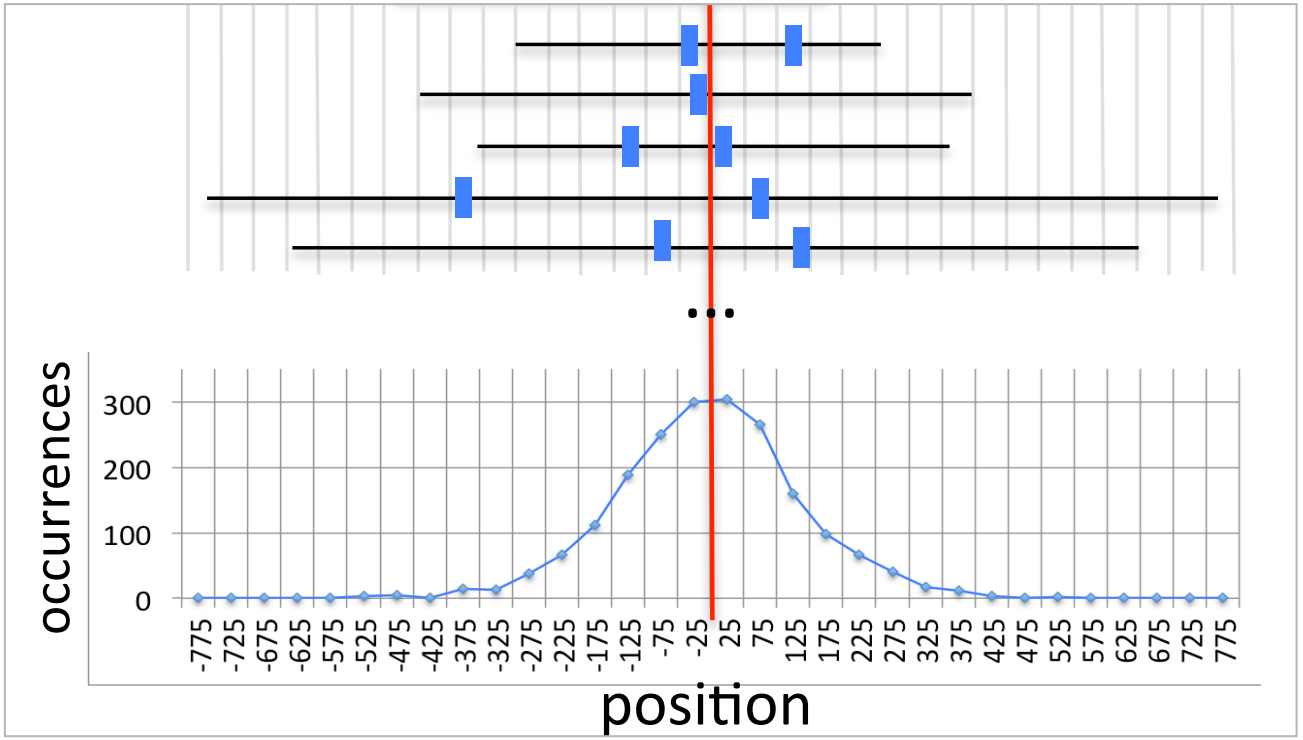


binding motif
(represented as a sequence logo)

- Adapted from Morgane Thomas-Chollier

Motif discovery ("de novo")

- Find exceptional motifs based on the sequence only
- (No prior knowledge of the motif to look for)
- Criteria of exceptionality:
  - ***Over-/under-representation:*** higher/lower frequency than expected by chance
  - ***Position bias:*** concentration at specific positions relative to some reference coordinates (e.g. TSS, peak center, …).



Slide Adapted from Morgane Thomas-Chollier. Drawing from Elodie Darbo

Some motif discovery tools

- Tools already exist for a long time !

- MEME (Bailey et al., 1994)
- **RSAT oligo-analysis (van Helden et al., 1998)**
- AlignACE (Roth et al. 1998)
- **RSAT position-analysis (van Helden et al., 2000)**
- Weeder (Pavesi et al. 2001)
- MotifSampler (Thijs et al., 2001)
- … many others

Regulatory sequence Analysis Tools (http://rsat.eu/)

Regulatory Sequence Analysis Tools

Welcome to **Regulatory Sequence Analysis Tools** (**RSAT**).

This web site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences.

RSAT servers have been up and running since 1997. The project was initiated by **Jacques van Helden**, and is now pursued by the **RSAT team**.

Choose a server

New ! *January 2015*: we are in the process of re-organising our mirror servers into taxon-specific servers, to better suit the drastic increase of available genomes.

maintained by TAGC - Université Aix Marseilles, France

maintained by RegulonDB - UNAM, Cuernavaca, Mexico

maintained by platforme ABIMS Roscoff, France

maintained by Ecole Normale Supérieure Paris, France

maintained by Bruno Contreras Moreira, Spain

maintained by SLU Global Bioinformatics Center, Uppsala, Sweden

Citing RSAT complete suite of tools:

- Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. (2011) **RSAT 2011: regulatory sequence analysis tools**. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W86-91. [Pubmed 21715389] [Full text]
- Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). **RSAT: regulatory sequence analysis tools**. Nucleic Acids Res. [Pubmed 18495751] [Full text]
- van Helden, J. (2003). **Regulatory sequence analysis tools**. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6. [Pubmed 12824373] [Full text] [pdf]

For citing individual tools: the reference of each tool is indicated on top of their query form.

RSAT logos designed by Mauricio Guzman (http://www.altamirastudio.com/)

RSAT supported genomes (Oct 2015)

Regulatory Sequence Analysis Tools

Welcome to **Regulatory Sequence Analysis Tools** (**RSAT**).

This web site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences.

RSAT servers have been up and running since 1997. The project was initiated by **Jacques van Helden**, and is now pursued by the **RSAT team**.

Choose a server

New ! *January 2015*: we are in the process of re-organising our mirror servers into taxon-specific servers, to better suit the drastic increase of available genomes.

maintained by TAGC - Université Aix Marseilles, France

maintained by RegulonDB - UNAM, Cuernavaca, Mexico

maintained by platforme ABIMS Roscoff, France

maintained by Ecole Normale Supérieure Paris, France

maintained by Bruno Contreras Moreira, Spain

maintained by SLU Global Bioinformatics Center, Uppsala, Sweden

Citing RSAT complete suite of tools:

- Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. (2011) **RSAT 2011: regulatory sequence analysis tools**. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W86-91. [Pubmed 21715389] [Full text]
- Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). **RSAT: regulatory sequence analysis tools**. Nucleic Acids Res. [Pubmed 18495751] [Full text]
- van Helden, J. (2003). **Regulatory sequence analysis tools**. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6. [Pubmed 12824373] [Full text] [pdf]

For citing individual tools: the reference of each tool is indicated on top of their query form.

RSAT logos designed by Mauricio Guzman (http://www.altamirastudio.com/)

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


Sylvain Brohée
Postdoc


Nicolas Simonis
Postdoc


Didier Croes
Postddoc


Didier Gonze
Premier assistant

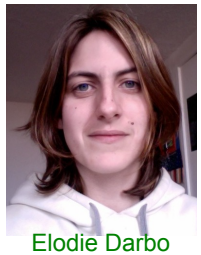

Myriam Loubriat
Secretary


Ariane Toussaint
Professor Emeritus


Jacques van Helden
Professor


Leon Juvenal Hajingabo
PhD Student


Maud Vidick
PhD Student (co-direction)


Elodie Darbo
PhD Student
co-direction Marseille


Alejandra Medina
PhD Student
co-direction Mexico


Morgane Thomas-Chollier
PhD student+postdoc


Matthieu Defrance
Postdoc


Olivier Sand
Postdoc


Jean Valéry Turatsinze
PhD student


Raphaël Leplae
Postdoc


Gipsi Lima
PhD + Postdoc


Karoline Faust
PhD student

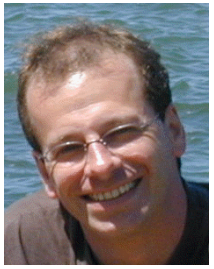

Rekin' s Janky
PhD student


Eric Vervisch
Research fellow


Guy Bottu
Postddoc

- **Conception, implementation, evaluation and application of bioinformatics methods for the analysis of genomes and biomolecular networks.**
- **Regulatory sequences**
  - Motif analysis algorithms *(Olivier Sand , Matthieu Defrance , Maud Vidick, Alejandra Medina-Rivera)*
  - Evolution of cis-acting elements in Bacteria (*Rekin's Janky, Alejandra Medina-Rivera*)
  - Regulation of development in Drosophila (*Jean Valéry Turatsinze, Elodie Darbo*)
  - Hox regulation in Vertebrates (*Morgane Thomas-Chollier)*
  - Work flows on transcriptional regulation (*Olivier Sand, Eric Vervisch)*
- **Biomolecular networks**
  - Network analysis tools (*Sylvain Brohée)*
  - Inference of metabolic pathways (*Karoline Faust, Didier Croes*)
  - Host-virus interaction networks (*Nicolas Simonis, Leon Juvénal Hagingambo)*
  - Analysis of regulatory networks (*Sylvain Brohée, Rekin's Janky*)
- **Mobile genetic elements in prokaryotes** *(Raphaël Leplae, Gipsi Lima, Ariane Toussaint)*
- **Modelling of dynamical systems** *(Didier Gonze)*
- **e-Learning for bioinformatics** *(Guy Bottu)*

Collaborators involved in the development of the Regulatory Sequence Analysis Tools



Bruno André
(ULB, Bruxelles, Belgium)
Initiation of the RSAT project.
Conception of oligo-analysis.
Analysis of yeast regulation.

ULB

Denis Thieffry
(ENS, Paris, France)
ChIP-seq tools +
regulatory networks.

Carl Herrmann
(TAGC, Marseille, France)
ChIP-seq analysis (peak-motifs, compare-matrices).

Elodie Darbo
(TAGC, Marseille, France)
Analysis of co-expression clusters + ChIP-seq data (transcription factors, chromatin marks).

Julio Collado-Vides
(CCG, Cuernavaca - Mexico)
Initiation of the RSAT project
Analysis of regulation
in bacterial genomes

CCG
Centro de Ciencias Genómicas

Alejandra Medina-Rivera
(CCG, Cuernavaca - Mexico)
Evaluation of matrix quality.
Phylogenetic footprints in Baceria.

CCG
Centro de Ciencias Genómicas

Lionel Spinelli
(TAGC, Marseille, France)
Development of peak-footprints.

Cei Abreu-Goodger
(Sanger Institute, Hinxton, UK)
Evaluation of matrix quality
on bacterial regulons.

wellcome trust sanger institute

Practical - RSAT quick tour

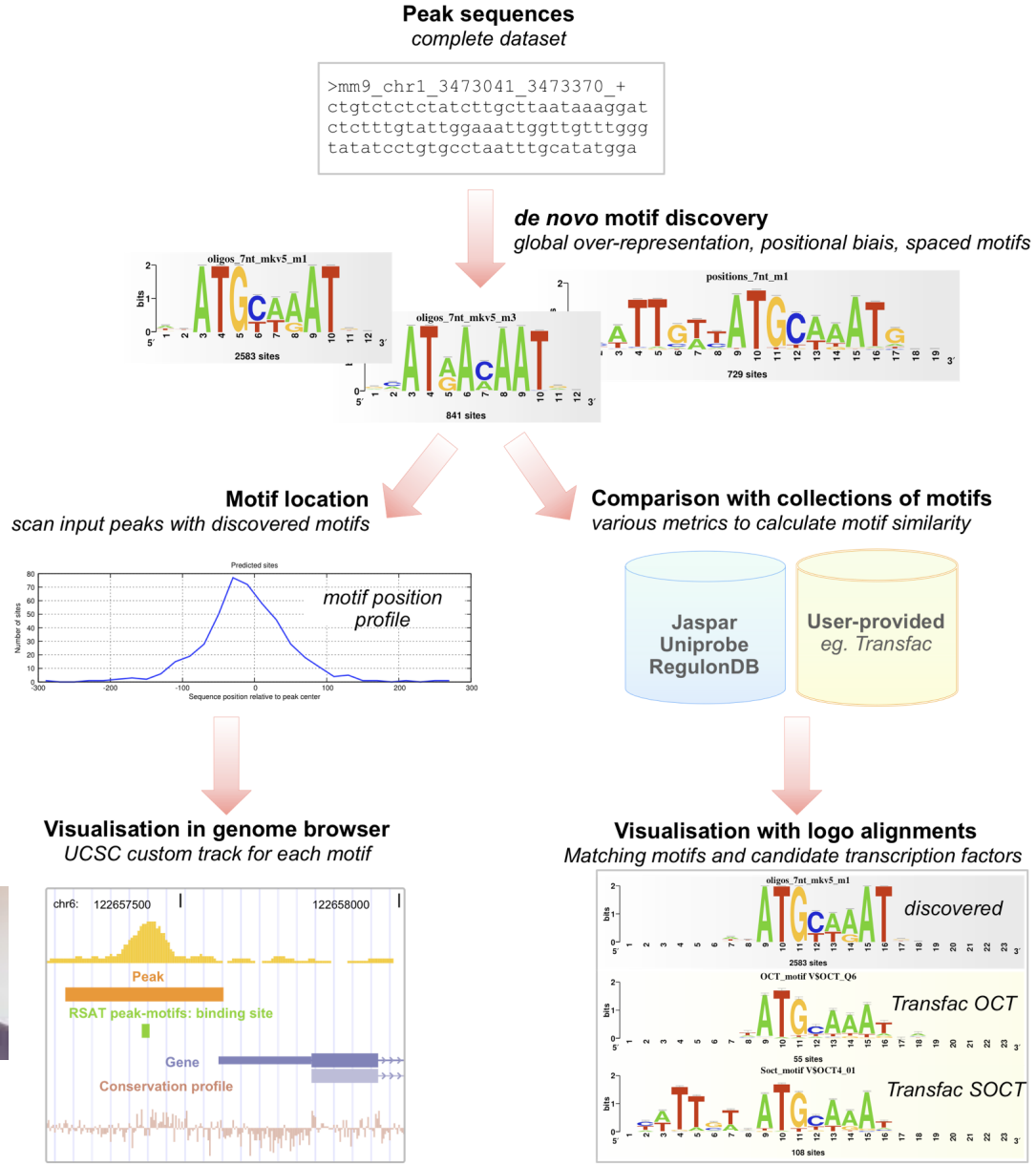
- Open a connection to the RSAT portal (http://rsat.eu/).
- Select the Metazoa server (hosted on the ABIMS platform at Roscoff - France).
- Explore the tools :
  - On the home page, use the 3-questions guidance and ***find a way to discover motifs in ChIP-seq peaks***.
  - In the left-side menu, click on the black boxes to expand thematic lists of tools, and browse the tool names to get a global idea of the supported functionalities.
  - In the ***Tutorials*** main page, inspect the tool map (RSAT flow chart) and think about a path that you could use to analyze to go from peak coordinates to discovered motifs.

More info: RSAT descriptions + protocols

- Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C., *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res*, **43**, W50–6.
- Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, **7**, 1551–1568.
- Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*, **40**, e31–e31.
- Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*, **39**, W86–91.
- Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res*, **36**, W119–27.
- Sand,O., Thomas-Chollier,M., Vervisch,E. and van Helden,J. (2008) Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. *Nature Protocols*, **3**, 1604–1615.
- Turatsinze,J.-V., Thomas-Chollier,M., Defrance,M. and van Helden,J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, **3**, 1578–1588.
- Defrance,M., Janky,R., Sand,O. and van Helden,J. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature Protocols*, **3**, 1589–1603.

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 fo discovered motifs in large sequence sets (tens 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 geno browsers (e.g. UCSC) for visualization.
- Interfaces
  - Stand-alone
  - Web interface
  - Web services (SOAP/WSDL)

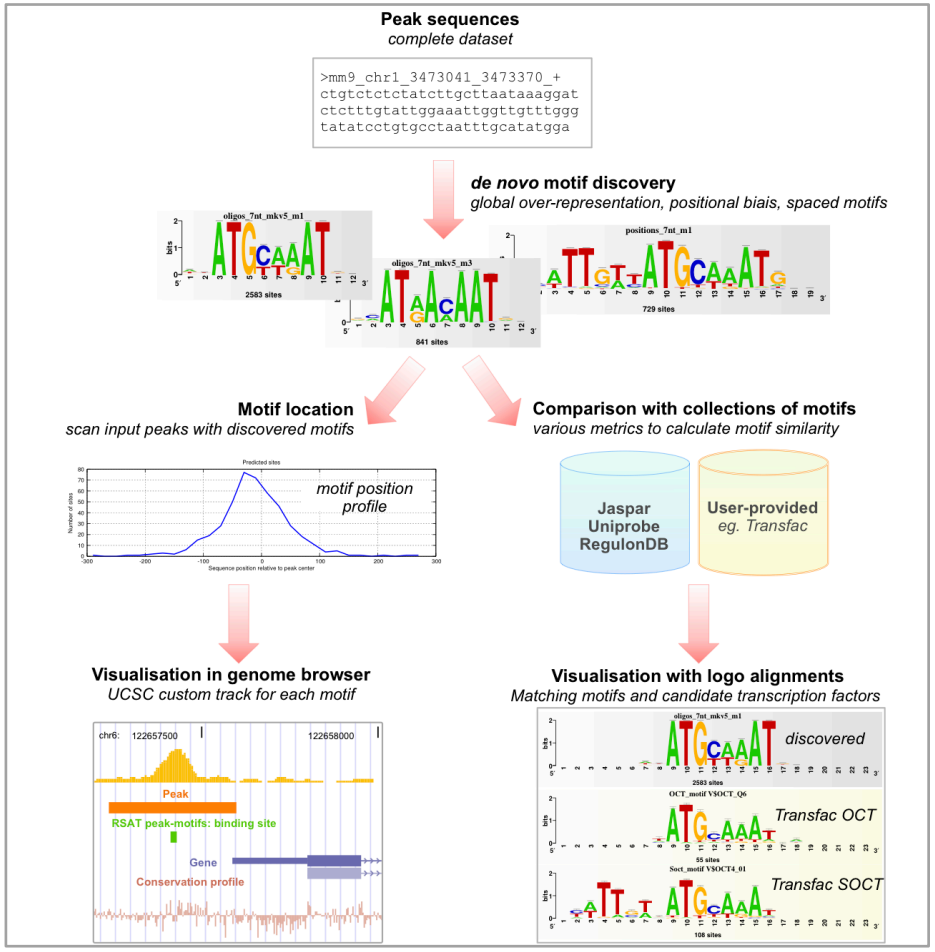


Morgane Thomas-Chollier | Matthieu Defrance | Olivier Sand | Carl Herrmann (TACG, Marseille) | Denis Thieffry ENS Paris

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.

New approaches for ChIP-seq datasets

- de novo motif discovery (peak-motifs in RSAT)



Thomas-Chollier et al Nucleic Acids Research, 2012

Peaks coordinates

BED

```
chr1    3002142 3002195
chr1    3002804 3002853
```

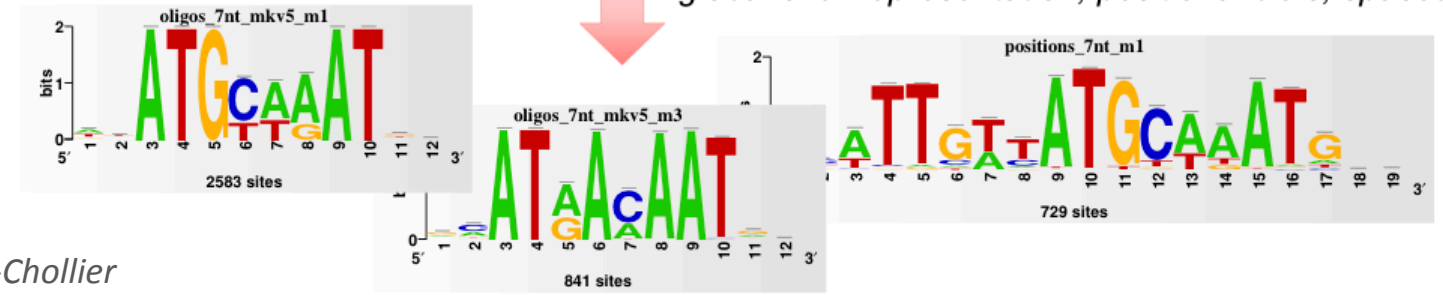
Peak sequences

complete dataset

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

***de novo* motif discovery**

global over-representation, positional biais, spaced motifs



Peaks coordinates

BED

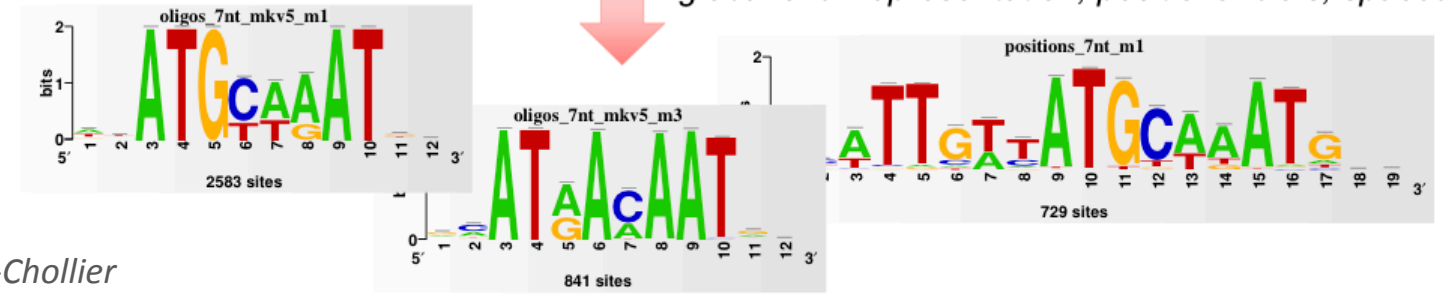
```
chr1    3002142 3002195
chr1    3002804 3002853
```

Extract corresponding sequences

Peak sequences
complete dataset

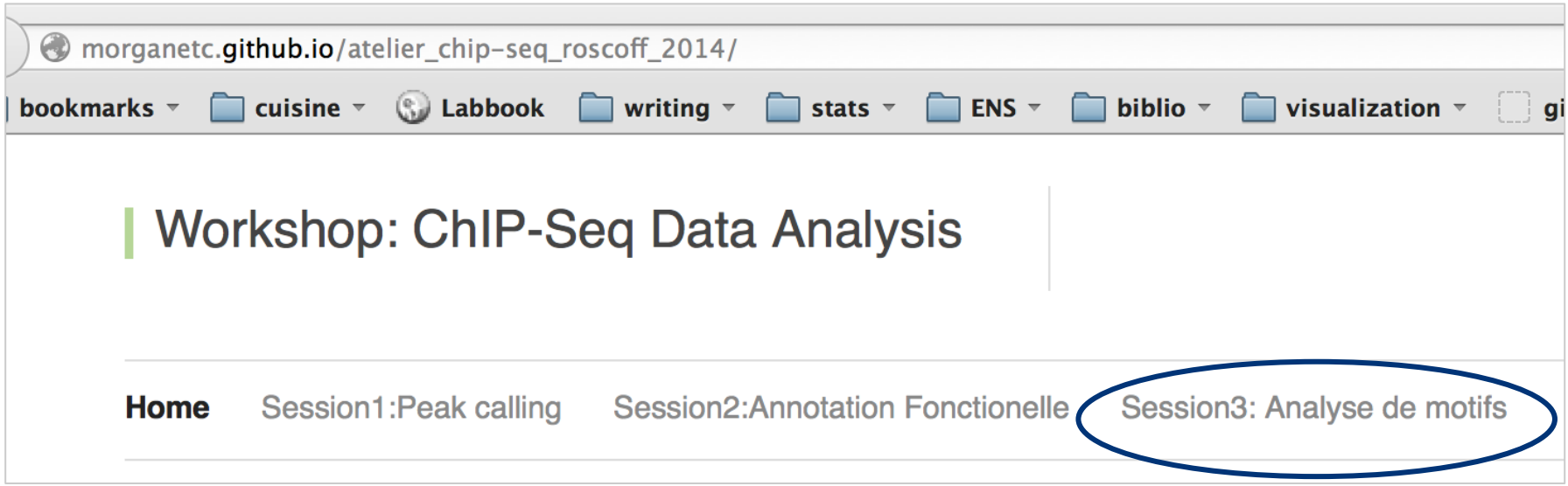
```
>mm9_chr1_3473041_3473370_+
ctgtctctctatcttgcttaataaaggat
ctctttgtattggaaattggttgtttggg
tatatcctgtgcctaatttgcatatgga
```

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

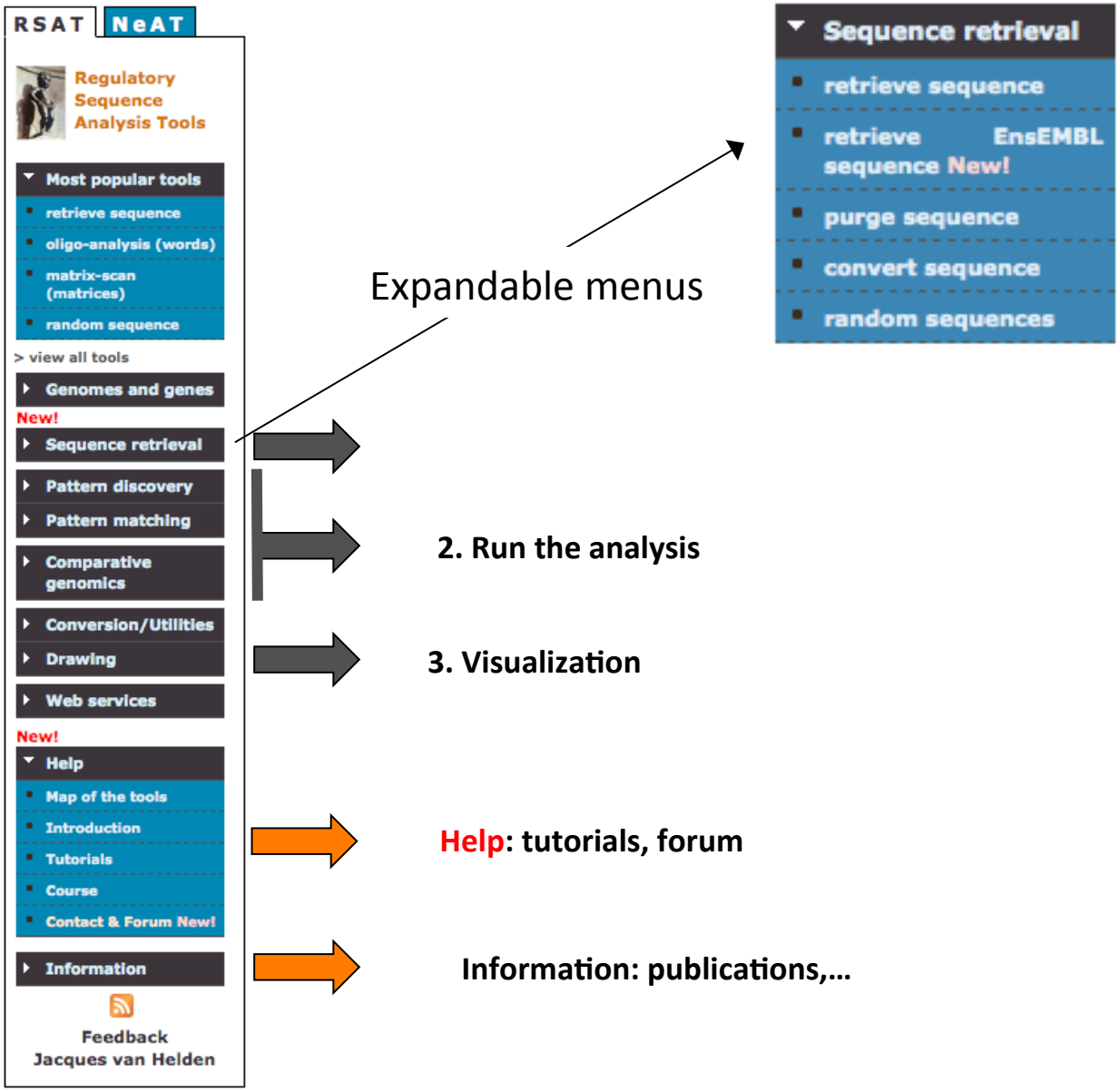


Hands on !

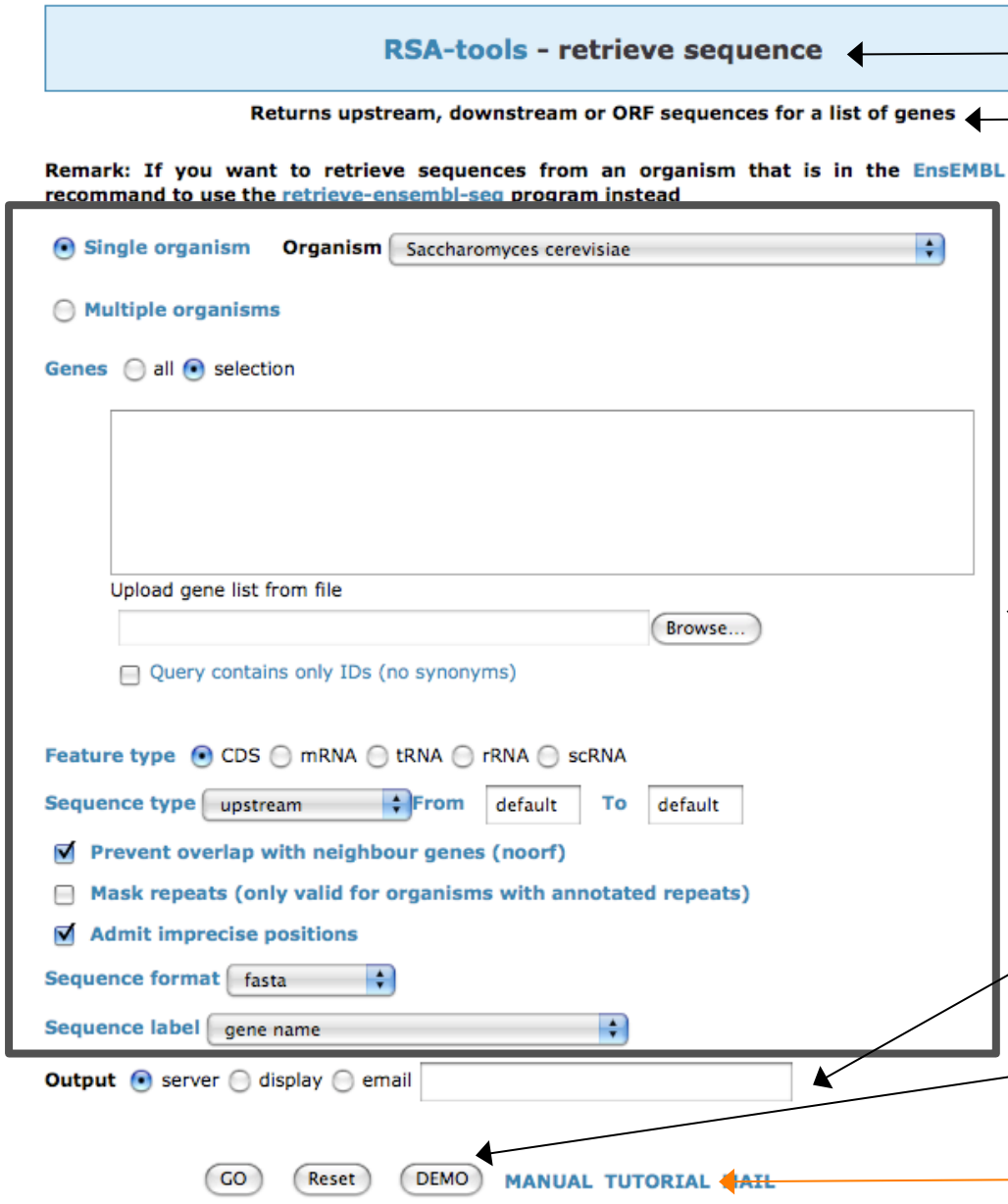
- Go to the companion website
  - http://ecole-bioinfo-aviesan.sb-roscoff.fr/
  - Click *Slides*
  - In the Section “ChIP-Seq”, follow the link *Tutorials: ChIP-seq analysis*
  - Open Session3: Analyse de motifs



Using RSAT

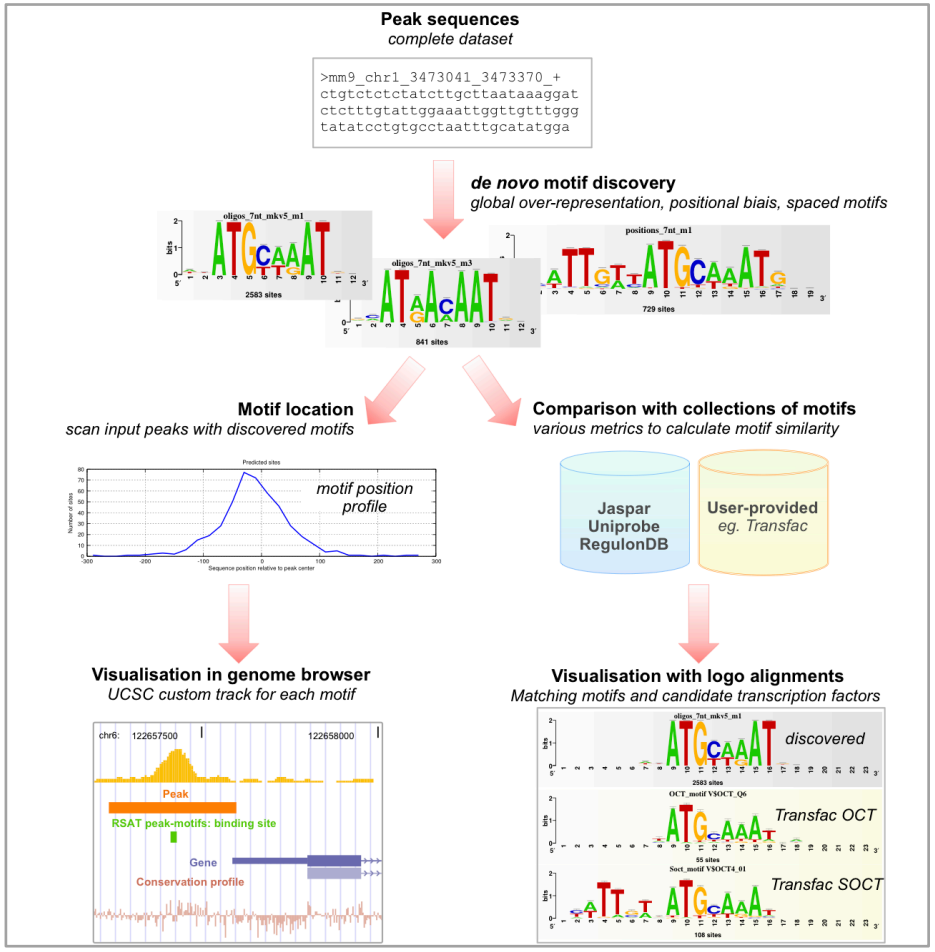

RSAT
NeAT
Regulatory Sequence Analysis Tools
Most popular tools
retrieve sequence
oligo-analysis (words)
matrix-scan (matrices)
random sequence
> view all tools
Genomes and genes
New!
Sequence retrieval
Pattern discovery
Pattern matching
Comparative genomics
Conversion/Utilities
Drawing
Web services
New!
Help
Map of the tools
Introduction
Tutorials
Course
Contact & Forum New!
Information
Feedback
Jacques van Helden
Expandable menus
Sequence retrieval
retrieve sequence
retrieve EnsEMBL sequence New!
purge sequence
convert sequence
random sequences
2. Run the analysis
3. Visualization
Help: tutorials, forum
Information: publications,...

RSAT Web forms



New approaches for ChIP-seq datasets

- de novo motif discovery (peak-motifs in RSAT)
- Note: "new" approaches are actually an adaptation of methods developed since 1998/2000 to treat huge data sets (server supports treatment of tens of Mb per analysis).



Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	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	-
Stand-alone version	yes	yes	no	yes	yes	yes
Tasks						
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	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	-
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*

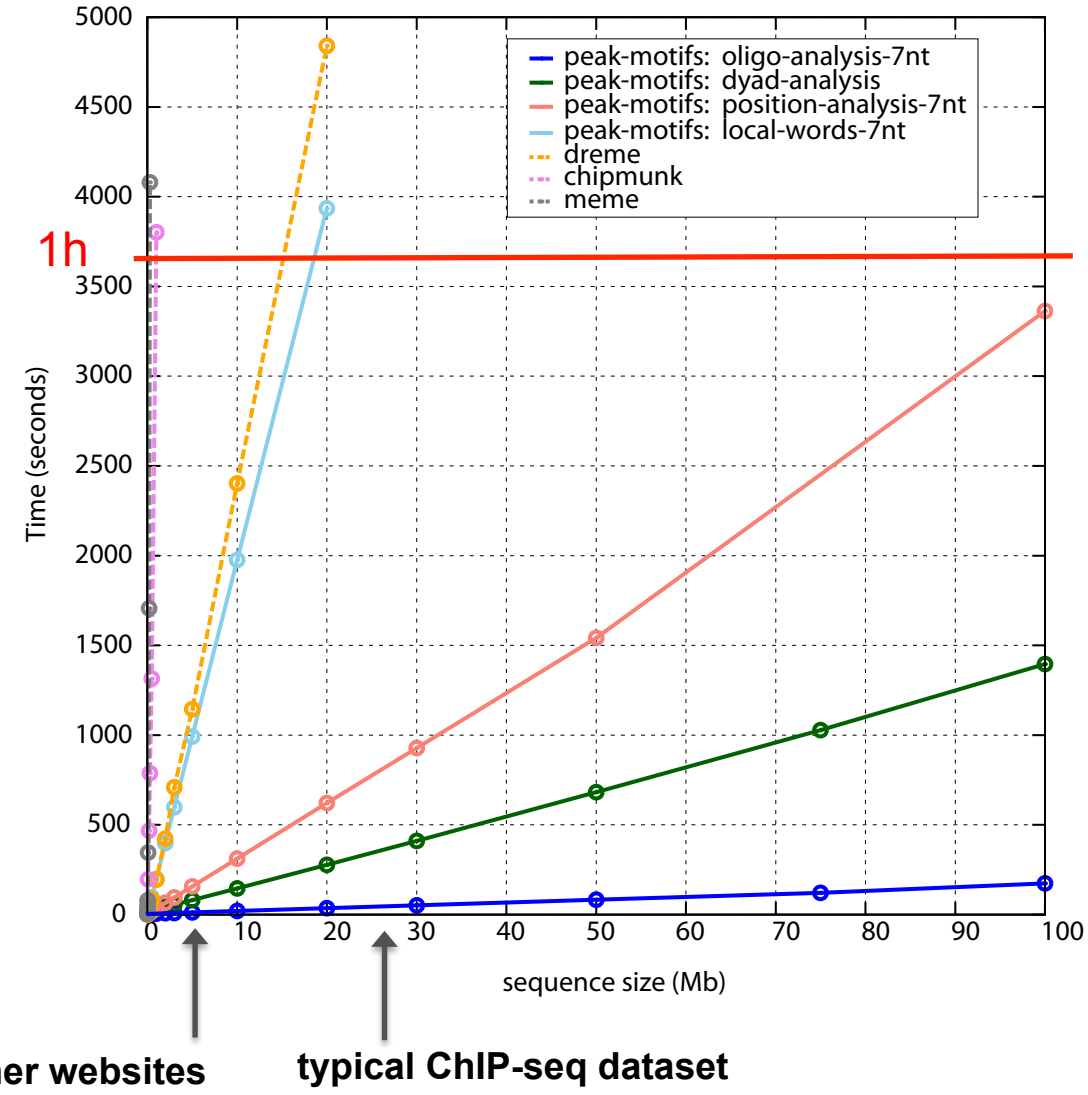
Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	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	-
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*

Peak-motifs: why providing yet another tool ?

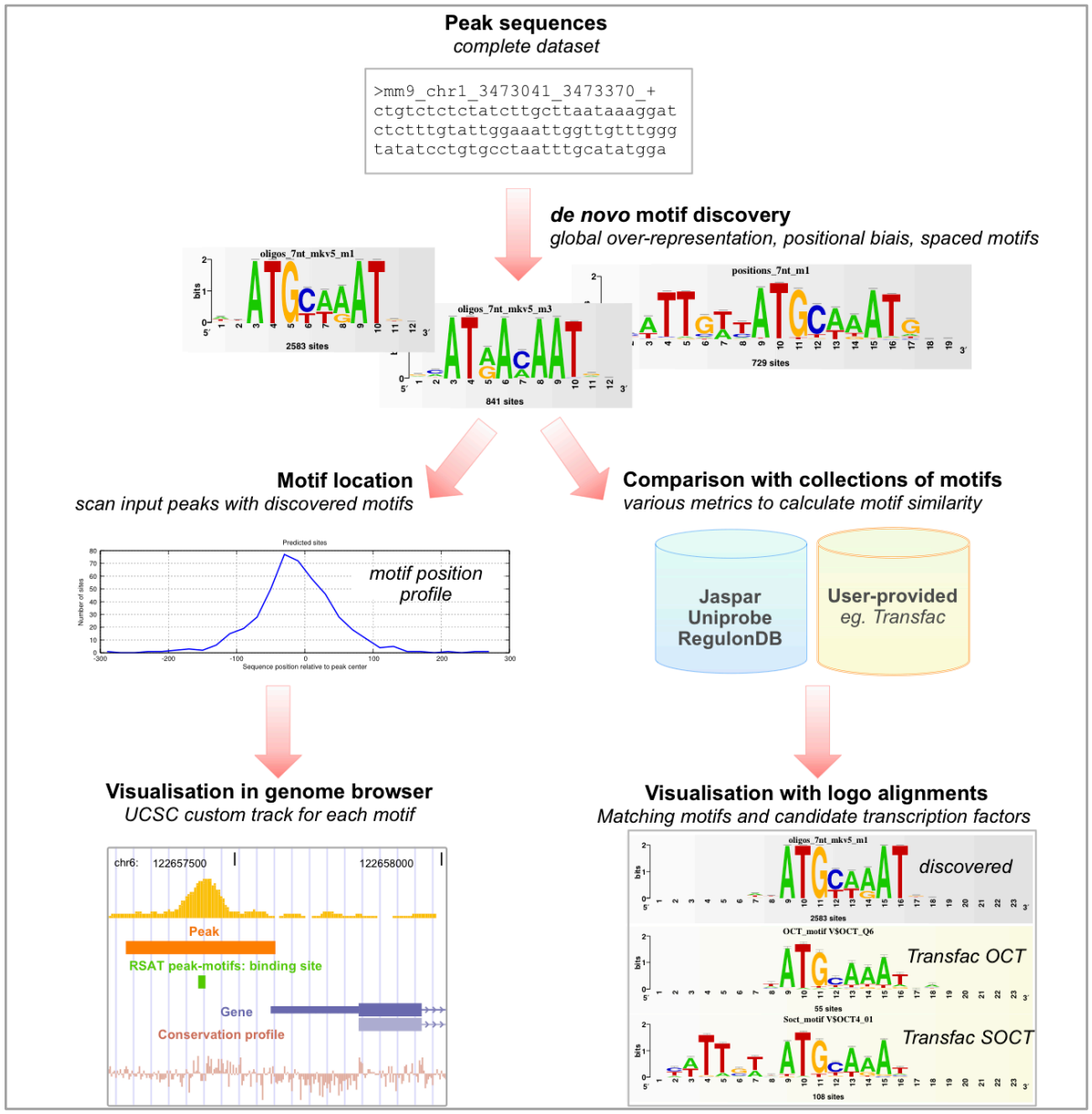
- **fast and scalable**
- **treat full-size datasets**



Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*

Peak-motifs: why providing yet another tool ?

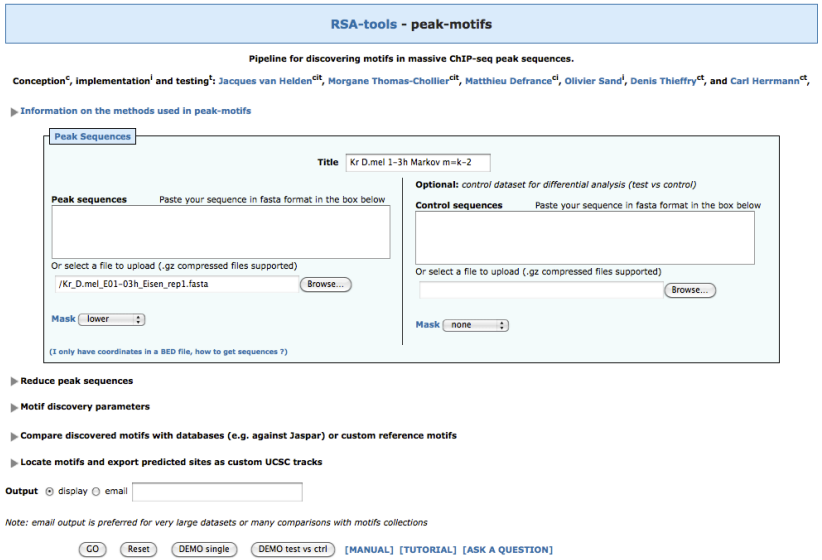
- **fast and scalable**
- **treat full-size datasets**
- **complete pipeline**
  - Peak properties (nucleotide, dinucleotide composition, lengths)
  - Motif discovery
  - Comparison with known motifs
  - Peak scanning



Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*

Peak-motifs: why providing yet another tool ?

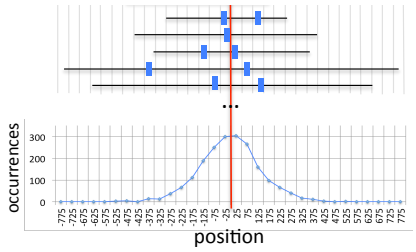
- **fast and scalable**
- **treat full-size datasets**
- **complete pipeline**
- **accessible to non-specialists**
  - Demo buttons
  - Tutorials & Protocols
    - Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden Nature Protocols,2012
  - Human-readable HTML report with links to all result files.

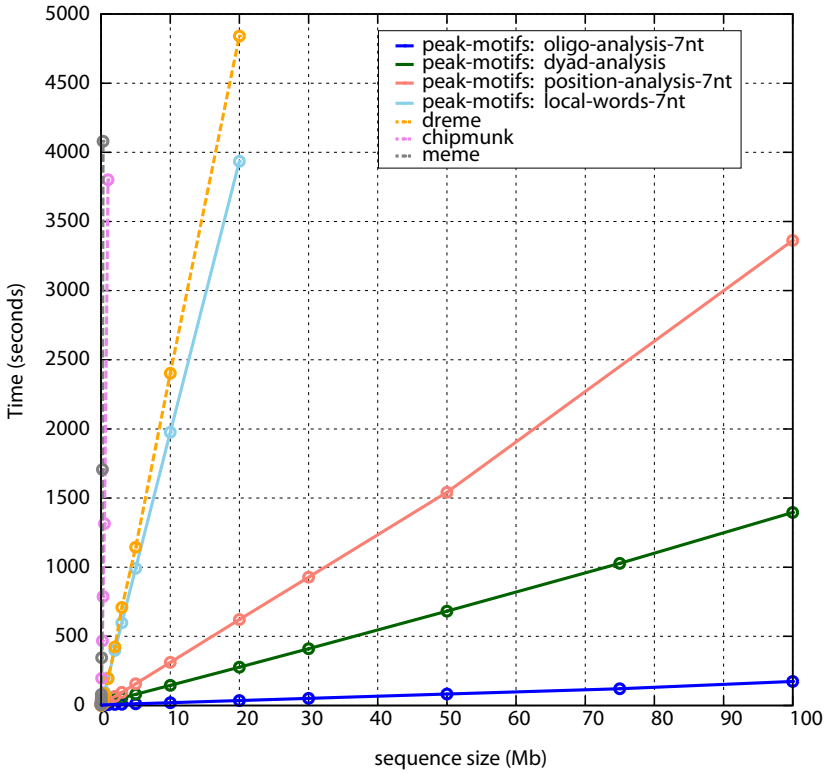


Adapted from Morgane Thomas-Chollier

Peak-motifs: why providing yet another tool ?

- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface
- accessible to non-specialists
- **using 4 complementary algorithms**
  - **Global over-representation**
    - **oligo-analysis**
    - **dyad-analysis (spaced motifs)**
  - **Positional bias**
    - **position-analysis**
    - **local-words**

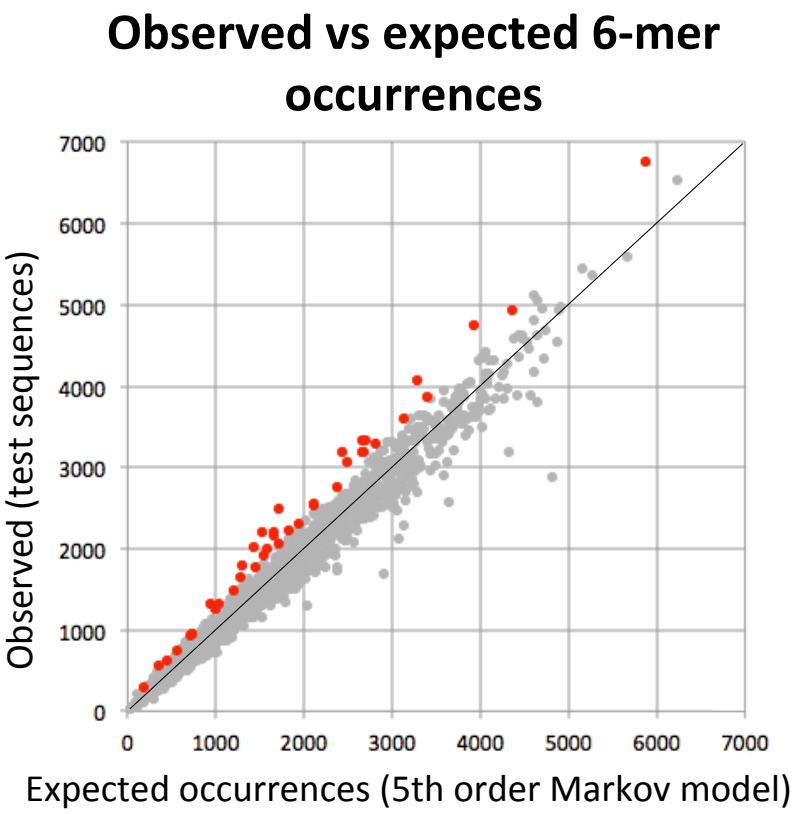




Motif discovery methods: frequency

Observed 6-mer occurrences computed from:	**Expected** 6-mer occurrences computed from:
6-mer (e.g.AACAAA)	**Background sequences (when available)**
Test sequences	OR
	Theoretical k-mers frequencies from test sequences

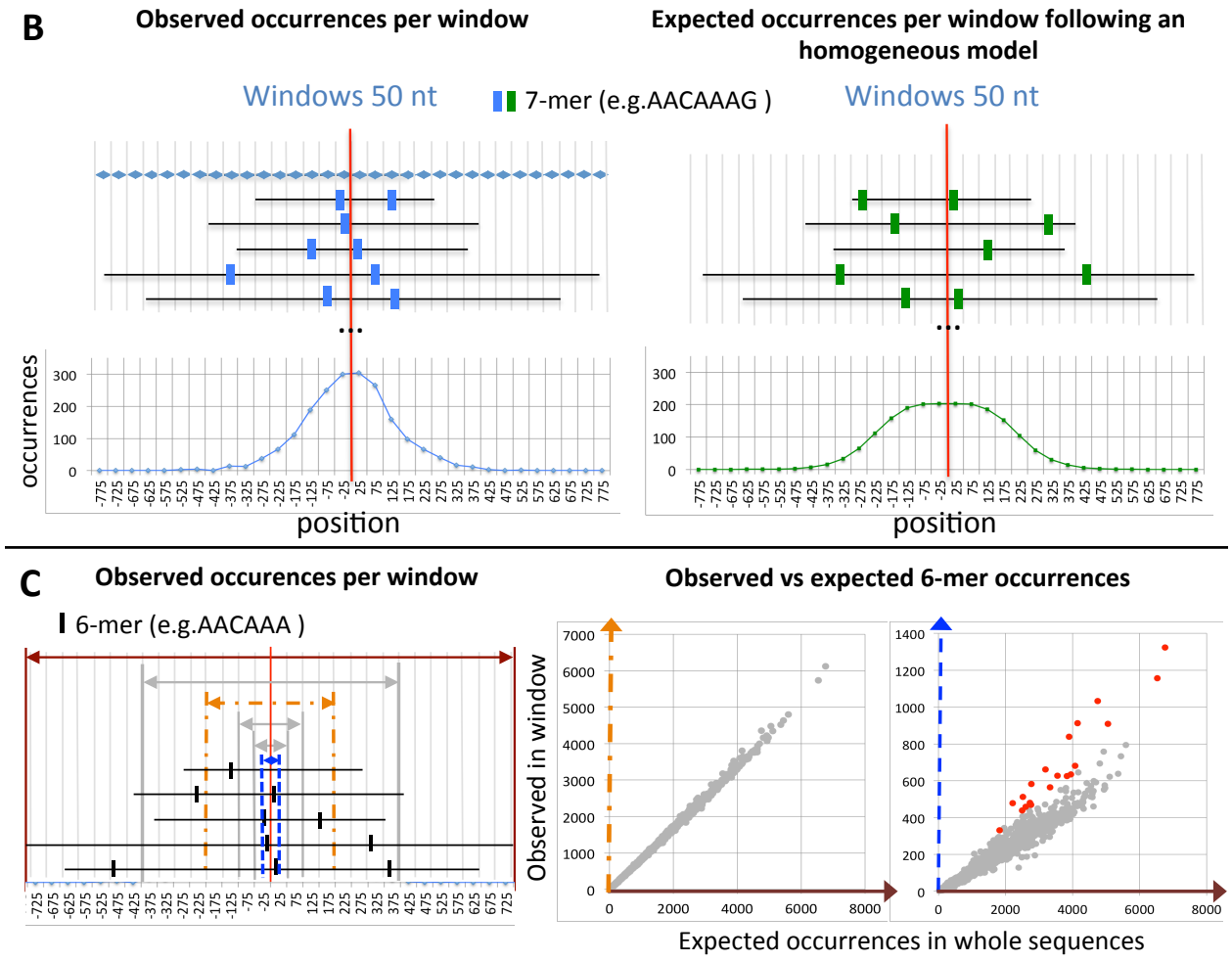
➔ Computation of p-value (binomial) and E-value (multi-testing correction)



oligo-analysis
dyad-analysis (spaced motifs)

Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van He

Motif discovery methods: positional bias



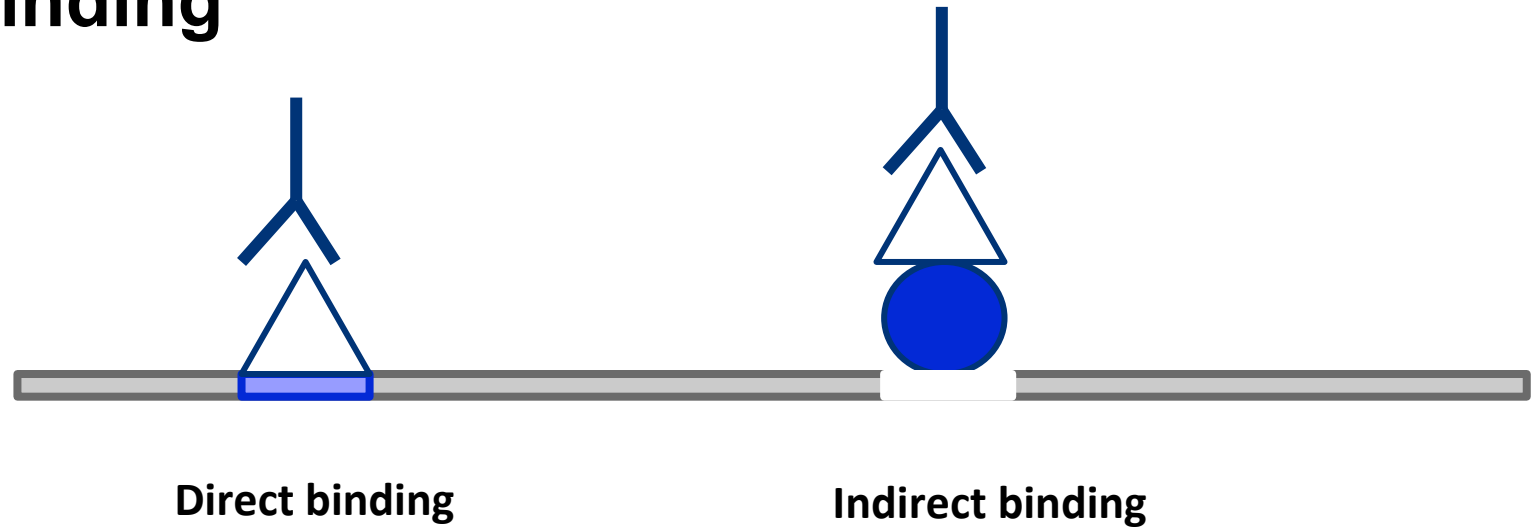
position-analysis

local-words

Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden ***Nature Protocols****,2012*

Direct versus indirect binding

- ChIP-seq does not necessarily reveal **direct binding**

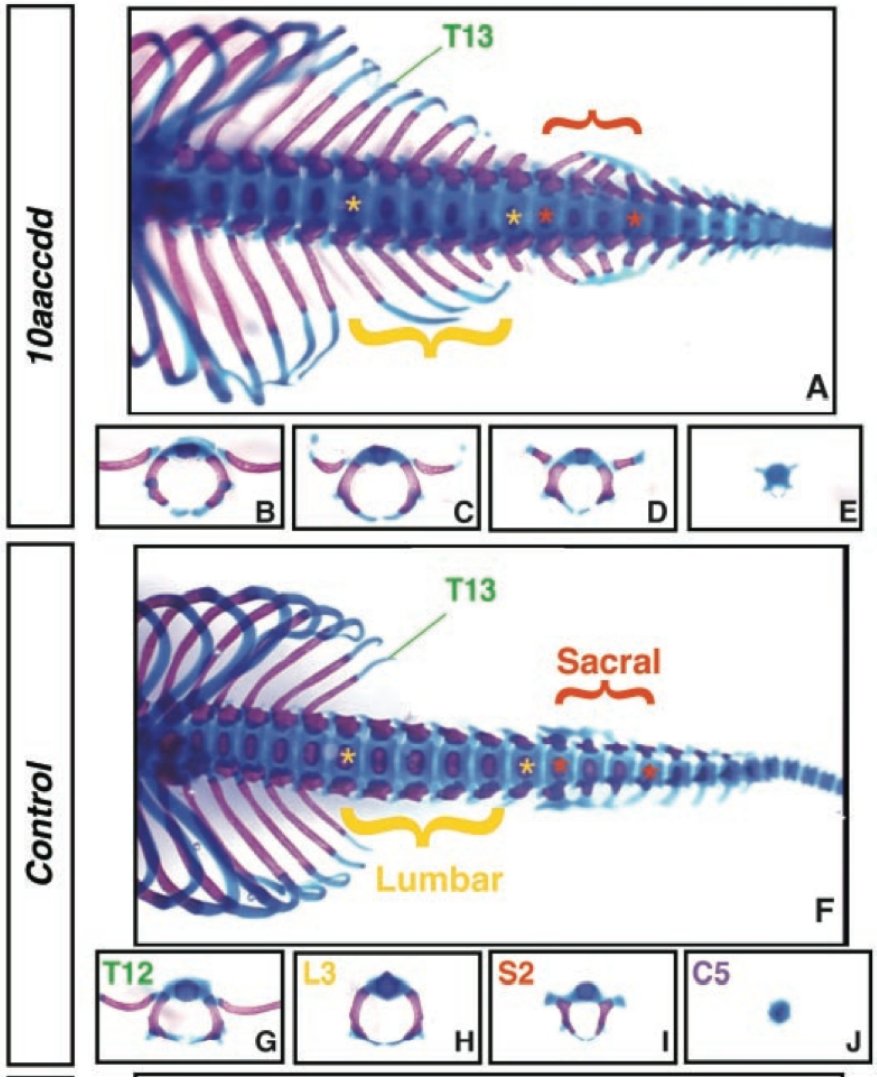


The motif of the targeted TF is not always found in peaks !

Hands on !

- Go to the companion website
  - http://dputhier.github.io/EBA_2015_ChIP-Seq/tutorial/04_motif/motif_tutorial.html
- Follow **steps 1 & 2** of **Discovering motifs from peak sequences**

Negative Controls in biology

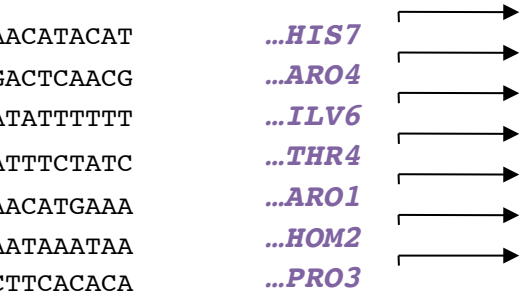


Wellik and Mario R Capecchi, Science, 2003

In the context of cis-regulation

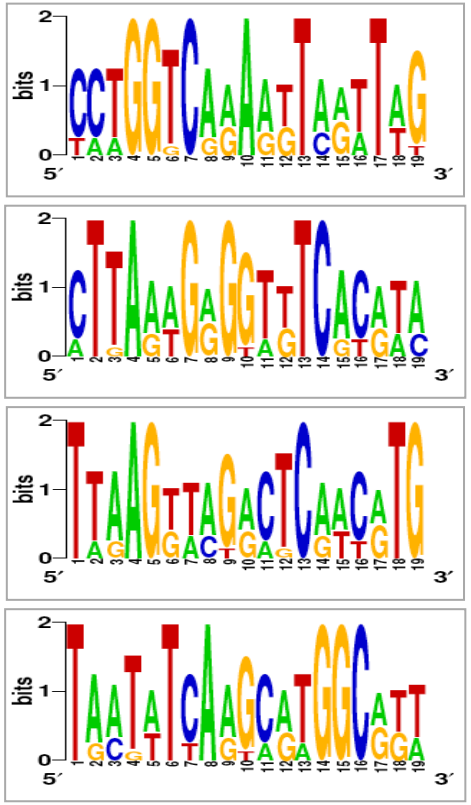
Use different set of ***sequences*** *(supposed to contain no exceptional motif)*

5′– TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT	…HIS7
5′– ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG	…ARO4
5′– CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT	…ILV6
5′– TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC	…THR4
5′– ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA	…ARO1
5′– ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA	…HOM2
5′– GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA	…PRO3



Use different set of *matrices*

(supposed to correspond to no specific factor)



Sequences

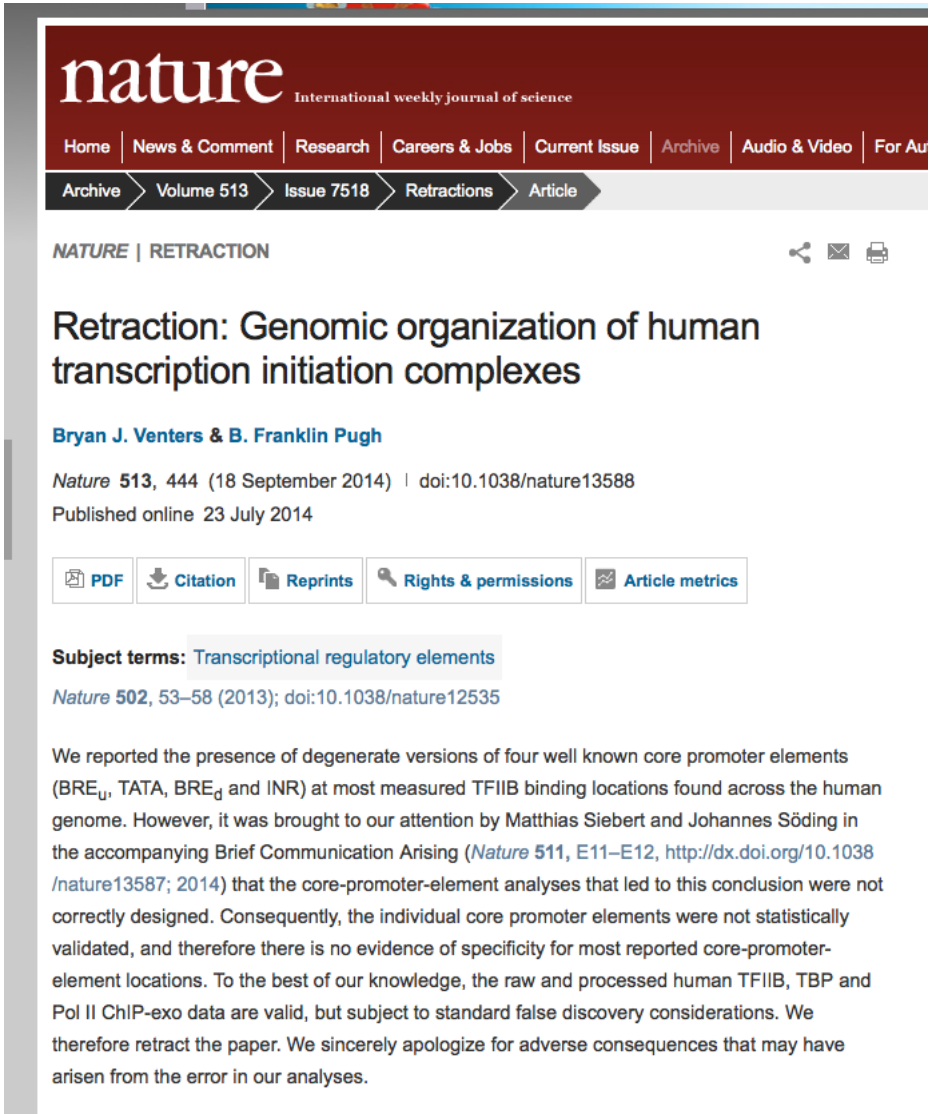
- **Positive control**: quantify the capability of the program to detect known regulatory elements
  - Annotated sites (e.g. sites from TRANSFAC) in their original context (the promoter sequences).
  - Annotated sites implanted in other context
    - Biological sequences (random selection).
    - Artificial sequences.
  - Artificial sites implanted in artificial sequences.

- **Negative control**: quantify the capability of the program to return a negative answer when there are no regulatory elements.
  - Artificial sequences
    (generated according to a Bernoulli or a Markov model to mimic an organism of interest)
  - Biological sequences without common regulation
    (random selection of genes)

Biological sequences

- Random genome fragments in RSAT
  - Select a set of fragments with random positions in a given genome, and return their coordinates and/or sequences
  - Adapted to chip-seq ?
    - Yes: same number of peaks + same size
    - No: composition of the sequences (dinucleotides) not respected
  - Complexify the control :
    - Make sure no peak is covered
    - Take regions close / far from the peaks
    - Maintain same composition
    - Maintain same dataset size
    - …

Why is it important ?



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive > Volume 513 > Issue 7518 > Retractions > Article

NATURE | RETRACTION

Retraction: Genomic organization of human transcription initiation complexes

Bryan J. Venters & B. Franklin Pugh

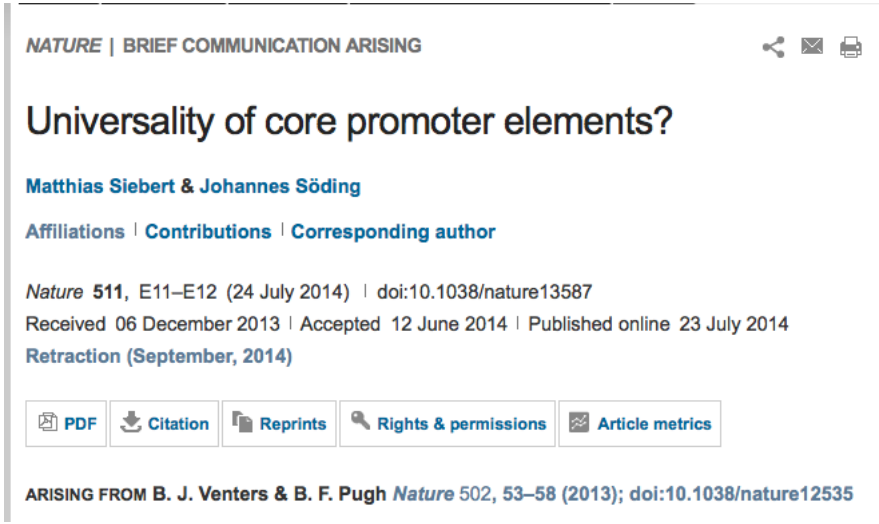
Nature **513**, 444 (18 September 2014) | doi:10.1038/nature13588
Published online 23 July 2014

PDF | Citation | Reprints | Rights & permissions | Article metrics

Subject terms: Transcriptional regulatory elements

Nature **502**, 53–58 (2013); doi:10.1038/nature12535

We reported the presence of degenerate versions of four well known core promoter elements (BRE_u, TATA, BRE_d and INR) at most measured TFIIB binding locations found across the human genome. However, it was brought to our attention by Matthias Siebert and Johannes Söding in the accompanying Brief Communication Arising (*Nature* **511**, E11–E12, http://dx.doi.org/10.1038/nature13587; 2014) that the core-promoter-element analyses that led to this conclusion were not correctly designed. Consequently, the individual core promoter elements were not statistically validated, and therefore there is no evidence of specificity for most reported core-promoter-element locations. To the best of our knowledge, the raw and processed human TFIIB, TBP and Pol II ChIP-exo data are valid, but subject to standard false discovery considerations. We therefore retract the paper. We sincerely apologize for adverse consequences that may have arisen from the error in our analyses.



NATURE | BRIEF COMMUNICATION ARISING

Universality of core promoter elements?

Matthias Siebert & Johannes Söding

Affiliations | Contributions | Corresponding author

Nature **511**, E11–E12 (24 July 2014) | doi:10.1038/nature13587
Received 06 December 2013 | Accepted 12 June 2014 | Published online 23 July 2014
Retraction (September, 2014)

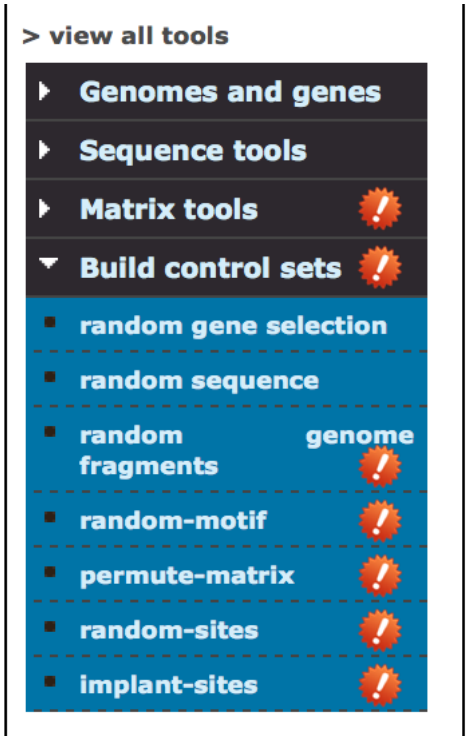
PDF | Citation | Reprints | Rights & permissions | Article metrics

ARISING FROM B. J. Venters & B. F. Pugh *Nature* 502, 53–58 (2013); doi:10.1038/nature12535

« We could reproduce one of the controls (60% GC random sequences) by assuming a wrong search space size of 1 instead of 161 (TATA), 60 (INR), or 40 (BRE_u and BRE_d), respectively. »

To prevent this ….

Building controls in RSAT



Hands on !

- Go to the companion website
  - http://dputhier.github.io/EBA_2015_ChIP-Seq/tutorial/04_motif/motif_tutorial.html
- Follow **step 3** of **Discovering motifs from peak sequences**

Hands on !

- Go to the companion website
  - http://dputhier.github.io/EBA_2015_ChIP-Seq/tutorial/04_motif/motif_tutorial.html
- Follow **step** of **Visualizing the sites in the context of genome annotations**

To go further

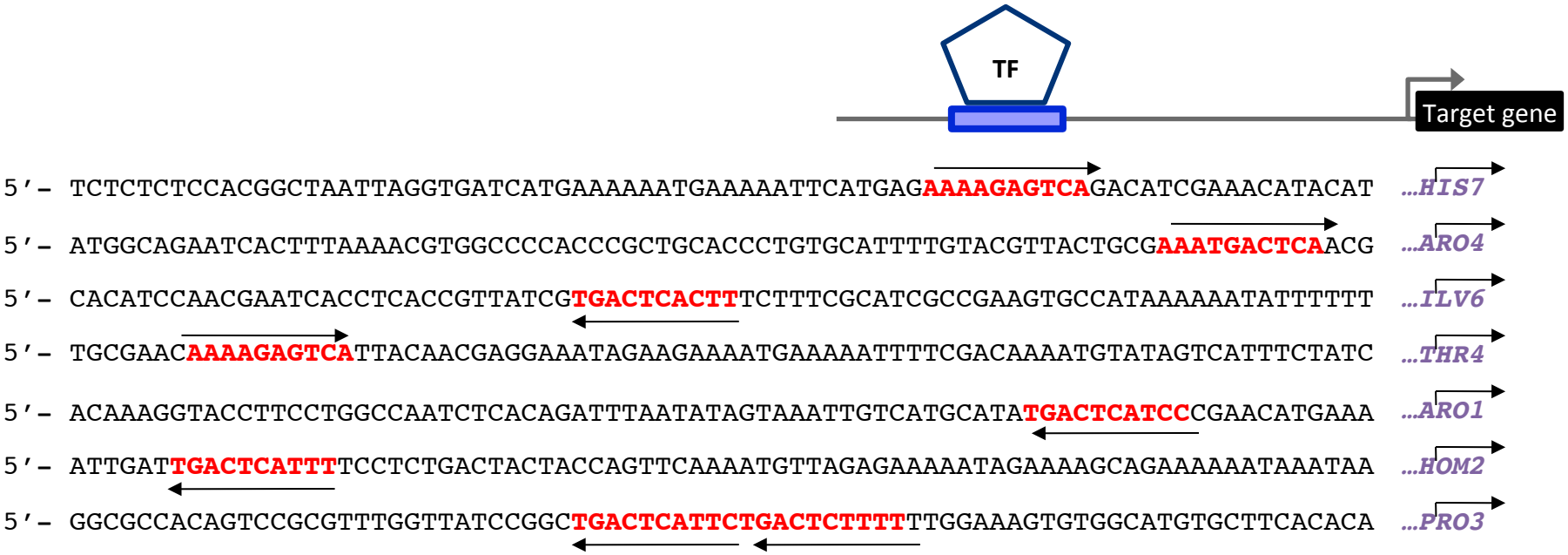
- The next slides explain step by step the algorithm behind oligo-analysis
- Peak-motifs : follow this protocol to grasp the detailed tweaking of parameters (send us an email to have free access to the PDF if necessary)
- Thomas-Chollier et al. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nature Protocols 7, 1551–1568 (2012).
- Matrix-quality : RSAT program that can be used to evaluate the enrichment of motifs in peaks
- Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J.Theoretical and empirical quality assessment of transcription factor-binding motifs.Nucleic Acids Res. 2011 Feb;39(3):808-24. doi: 10.1093/nar/gkq710. Epub 2010 Oct 4.

To go further

- Tutorial for ECCB 2014 : http://rsat.ulb.ac.be/eccb14/

- Master classes in analysis of cis-regulatory regions (over one week) at Ecole Normale Supérieure every september (contact : mthomas@biologie.ens.fr)

Principle: detect unexpected patterns



- Binding sites are represented as “words” = “string”=“k-mer”
  - e.g. **acgtga** is a 6-mer
- Signal is likely to be more frequent in the upstream regions of the co-regulated genes than in a random selection of genes
- We will thus detect over-represented words

Motif discovery using word counting

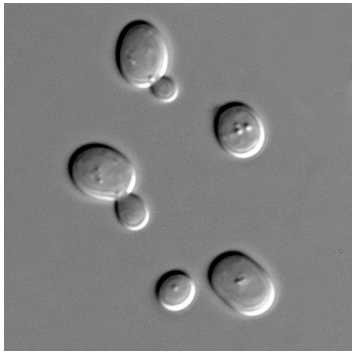
Idea:

motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)

Let's take an example (yeast Saccharomyces cerevisiae)

- NIT
  - 7 genes expressed under low nitrogen conditions
- MET
  - 10 genes expressed in absence of methionine
- PHO
  - 5 genes expressed under phosphate stress



	PHO
aaaaaa|tttttt	51
aaaaag|ctttttt	15
aagaaa|tttctt	14
gaaaaa|tttttc	13
tgccaa|ttggca	12
aaaaat|attttt	12
aaatta|taattt	12
agaaaa|ttttct	11
caagaa|ttcttg	11
aaacgt|acgttt	11
aaagaa|ttcttt	11
acgtgc|gcacgt	**10**
aataat|attatt	10
aagaag|cttctt	10
atataa|ttatat	10

	MET
aaaaaa|tttttt	105
atatat|atatat	41
gaaaaa|tttttc	40
tatata|tatata	40
aaaaat|attttt	35
aagaaa|tttctt	29
agaaaa|ttttct	28
aaaata|tatttt	26
aaaaag|ctttttt	25
agaaat|atttct	24
aaataa|ttattt	22
taaaaa|ttttta	21
tgaaaa|ttttca	21
ataata|tattat	20
atataa|ttatat	20

	NIT
aaaaaa|tttttt	80
cttatc|gataag	**26**
tatata|tatata	22
ataaga|tcttat	20
aagaaa|tttctt	20
gaaaaa|tttttc	19
atatat|atatat	19
agataa|ttatct	17
agaaaa|ttttct	17
aaagaa|ttcttt	16
aaaaca|tgtttt	16
aaaaag|ctttttt	15
agaaga|tcttct	14
tgataa|ttatca	14
atataa|ttatat	14

The most frequent oligonucleotides are not informative

- A (too) simple approach would consist in **detecting the most frequent oligonucleotides** (for example hexanucleotides) for each group of upstream sequences.
- This would however lead to deceiving results.
  - In all the sequence sets, the same kind of patterns are selected: **AT-rich hexanucleotides**.

	PHO
aaaaaa|tttttt	**51**
aaaaag|cttttt	15
aagaaa|tttctt	14
gaaaaa|ttttтc	13
tgccaa|ttggca	12
aaaaat|attttt	12
aaatta|taattt	12
agaaaa|ttttct	11
caagaa|ttcttg	11
aaacgt|acgttt	11
aaagaa|ttcttt	11
acgtgc|gcacgt	10
aataat|attatt	10
aagaag|cttctt	10
atataa|ttatat	10

	MET
aaaaaa|tttttt	**105**
atatat|atatat	41
gaaaaa|ttttтc	40
tatata|tatata	40
aaaaat|attttt	35
aagaaa|tttctt	29
agaaaa|ttttct	28
aaaata|tatttt	26
aaaaag|cttttt	25
agaaat|atttct	24
aaataa|ttattt	22
taaaaa|ttttta	21
tgaaaa|ttttca	21
ataata|tattat	20
atataa|ttatat	20

	NIT
aaaaaa|tttttt	80
cttatc|gataag	26
tatata|tatata	22
ataaga|tcttat	20
aagaaa|tttctt	20
gaaaaa|ttttтc	19
atatat|atatat	19
agataa|ttatct	17
agaaaa|ttttct	17
aaagaa|ttcttt	16
aaaaca|tgtttt	16
aaaaag|cttttt	15
agaaga|tcttct	14
tgataa|ttatca	14
atataa|ttatat	14

A more relevant criterion for over-representation

- The most frequent patterns do not reveal the motifs specifically bound by specific transcription factors.

- They merely **reflect the compositional biases** of upstream sequences.

- A more relevant criterion for over-representation is to detect patterns which **are more frequent** in the upstream sequences of the selected genes (co-regulated) **than the random expectation**.

- The **random expectation** is calculated by counting the frequency of each pattern in the complete set of upstream sequences (all genes of the genome).
  => **“Background”**

Motif discovery using word counting

Idea:

motifs corresponding to binding sites are generally repeated in the dataset

→ capture this statistical signal

- Algorithm
  - count occurrences of all k-mers in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the expected number of occurrences from a background model
    - empirical based on observed k-mer frequencies
    - theoretical background model (Markov Models)

Estimation of word expected frequencies from background sequences

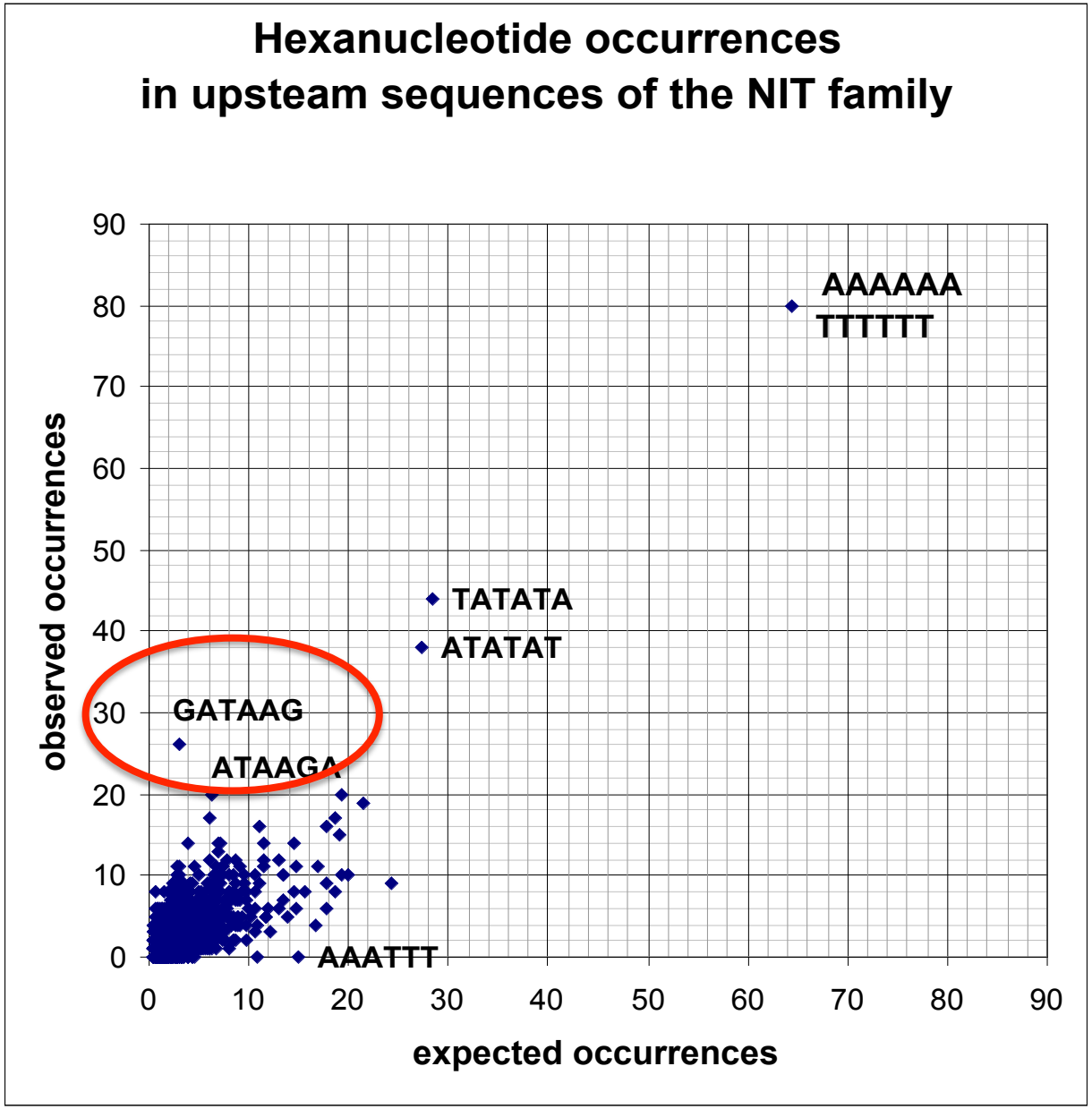
Example:

6nt frequencies in the whole set of 6000 yeast **upstream** sequences

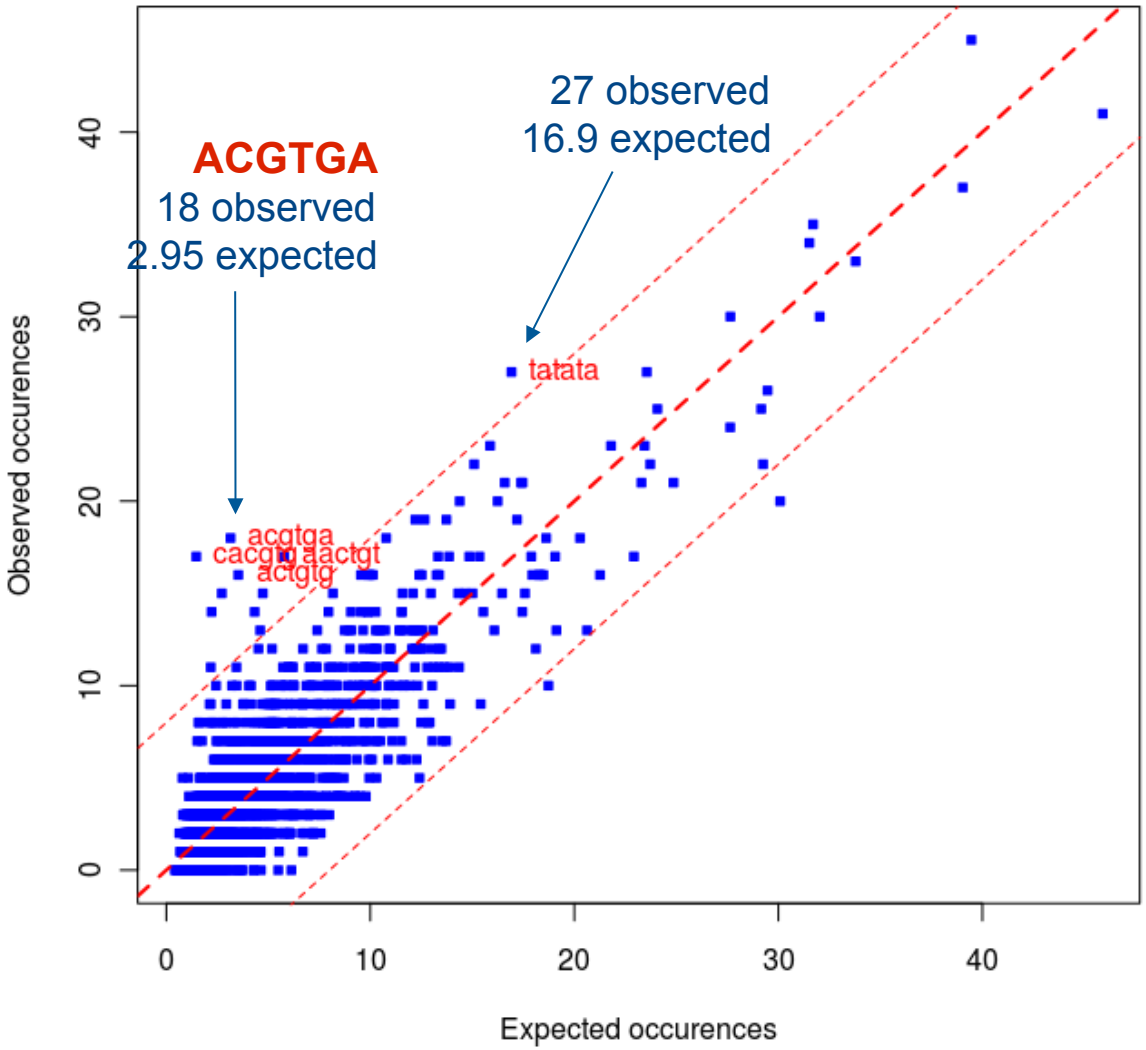
;seq	identifier	observed_freq	occ
aaaaaa	aaaaaa|ttttt	0,00510699	14555
aaaaac	aaaaac|gtttt	0,00207402	5911
aaaaag	aaaaag|ctttt	0,00375191	10693
aaaaat	aaaaat|atttt	0,00423577	12072
aaaaca	aaaaca|tgttt	0,0019828	5651
aaaacc	aaaacc|ggttt	0,00088526	2523
aaaacg	aaaacg|cgttt	0,00090105	2568
aaaact	aaaact|agttt	0,0014621	4167
aaaaga	aaaaga|tcttt	0,00323016	9206
aaaagc	aaaagc|gcttt	0,00135824	3871
aaaagg	aaaagg|ccttt	0,0017849	5087
aaaagt	aaaagt|acttt	0,0019035	5425
aaaata	aaaata|tattt	0,00336805	9599
aaaatc	aaaatc|gattt	0,00131368	3744
aaaatg	aaaatg|cattt	0,00185648	5291
aaaatt	aaaatt|aattt	0,00269156	7671
aaacaa	aaacaa|ttgtt	0,00209999	5985
aaacac	aaacac|gtgtt	0,00071684	2043
aaacag	aaacag|ctgtt	0,00096491	2750
aaacat	aaacat|atgtt	0,00108982	3106
aaacca	aaacca|tggtt	0,00074421	2121

Hexanucleotide occurrences in yeast Nitrogen-responding genes

	NIT
aaaaaa|tttttt	80
cttatc|gataag	**26**
tatata|tatata	22
ataaga|tcttat	**20**
aagaaa|tttctt	20
gaaaaa|tttttc	19
atatat|atatat	19
agataa|ttatct	17
agaaaa|ttttct	17
aaagaa|ttcttt	16
aaaaca|tgtttt	16
aaaaag|cttttt	15
agaaga|tcttct	14
tgataa|ttatca	14
atataa|ttatat	14



Motif discovery using word counting



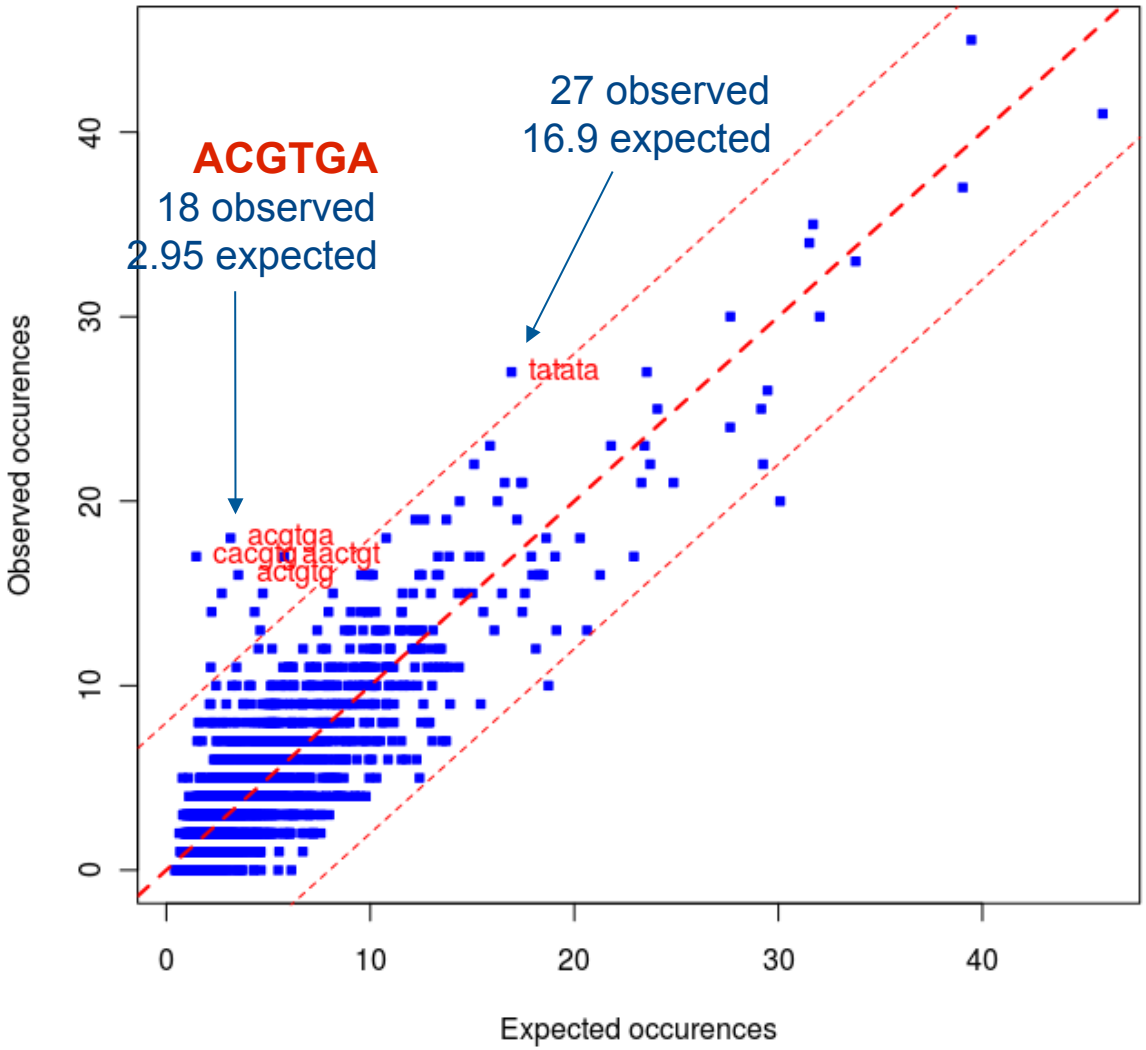
How to evaluate expected number of occurrences ?

Motif discovery using word counting

Idea:
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the **expected number of occurrences** from a background model
    - empirical based on observed k-mer frequencies
    - theoretical background model (Markov Models)
  - **statistical evaluation of the deviation observed** (P-value/E-value)

Statistical significance



How « big » is the surprise to observe 18 occurrences when we expect 2.95 ?

Statistical significance

How « big » is the surprise to observe 18 occurrences when we expect 2.95 ?

- at each position in the sequence, there is a **probability *p*** that the word starting at this position is ACGTGA
- we consider ***n*** positions
- what is the probability that ***k*** of these ***n*** positions correspond to ACGTGA ?
- **Application** : p = 3.4e-4 (intergenic frequencies)
  n = 9000 position
  x = 18 observed occurences

$$P(X \geq x) = \sum_{i=x}^{T} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

Binomial distribution to measure the "surprise"