

Détection de motifs dans les pics ChIP-seq

**M. Thomas-Chollier, M. Defrance,
C. Herrmann, D. Puthier**

Goal and organisation of this session

Goal: introduction to motif analysis in ChIP-seq data

- **processing steps:** from reads to peaks. => see previous session
- **downstream analyses:**
 - focus on motif analyses

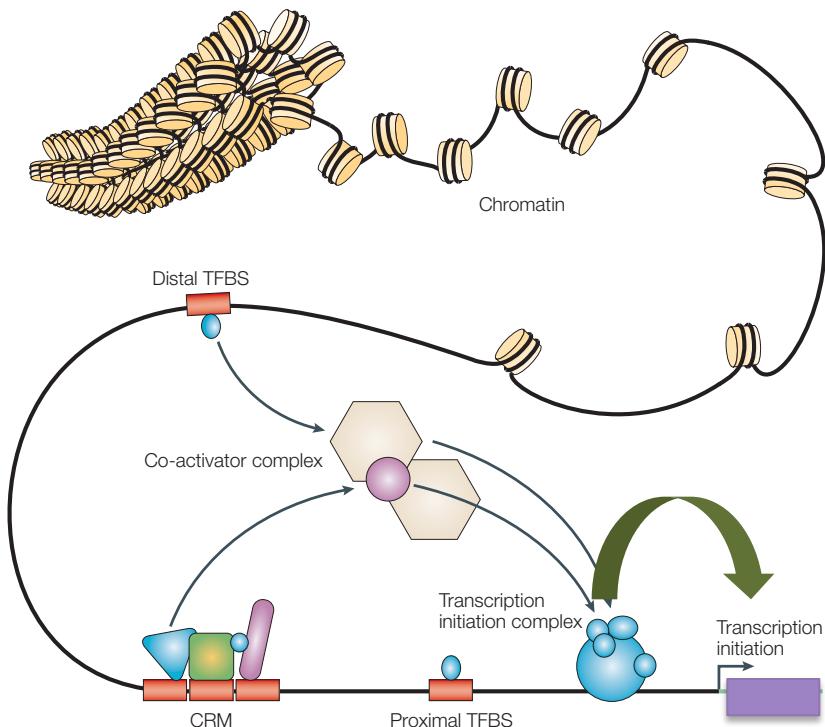
Tools

- RSAT

Steps

1. Retrieving **sequences** from a set of peak coordinates (*fetch-sequences*)
2. Discovering **motifs** from peak sequences (*peak-motifs*)
3. **Visualizing** the sites in the context of genome annotations (*UCSC genome browser*)

Biological concepts of transcriptional regulation



Wasserman et al, Nat Rev Genet, 2004

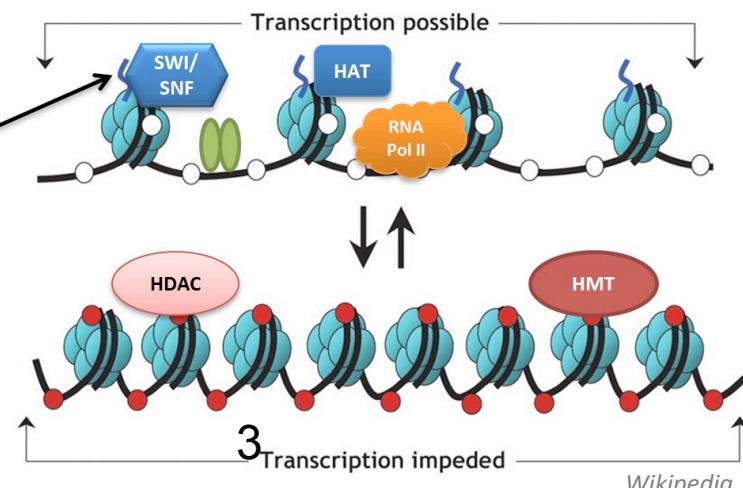
Chromatin accessibility (open/close) and **histone modifications** (eg: acetylation) also regulate gene expression

Transcription factors are proteins that modulate (activate/repress) the expression of **target genes** through the binding on **DNA cis-regulatory elements**

- Gene “switched on”
- Active (open) chromatin
 - Unmethylated cytosines (white circles)
 - Acetylated histones

(●) Transcription Factors / Co-activators

- Gene “switched off”
- Silent (condensed) chromatin
 - Methylated cytosines (red circles)
 - Deacetylated histones



in vivo experimental methods to identify binding sites

ChIP (=Chromatin Immuno-Precipitation)

=> differences in methods
to detect the bound DNA

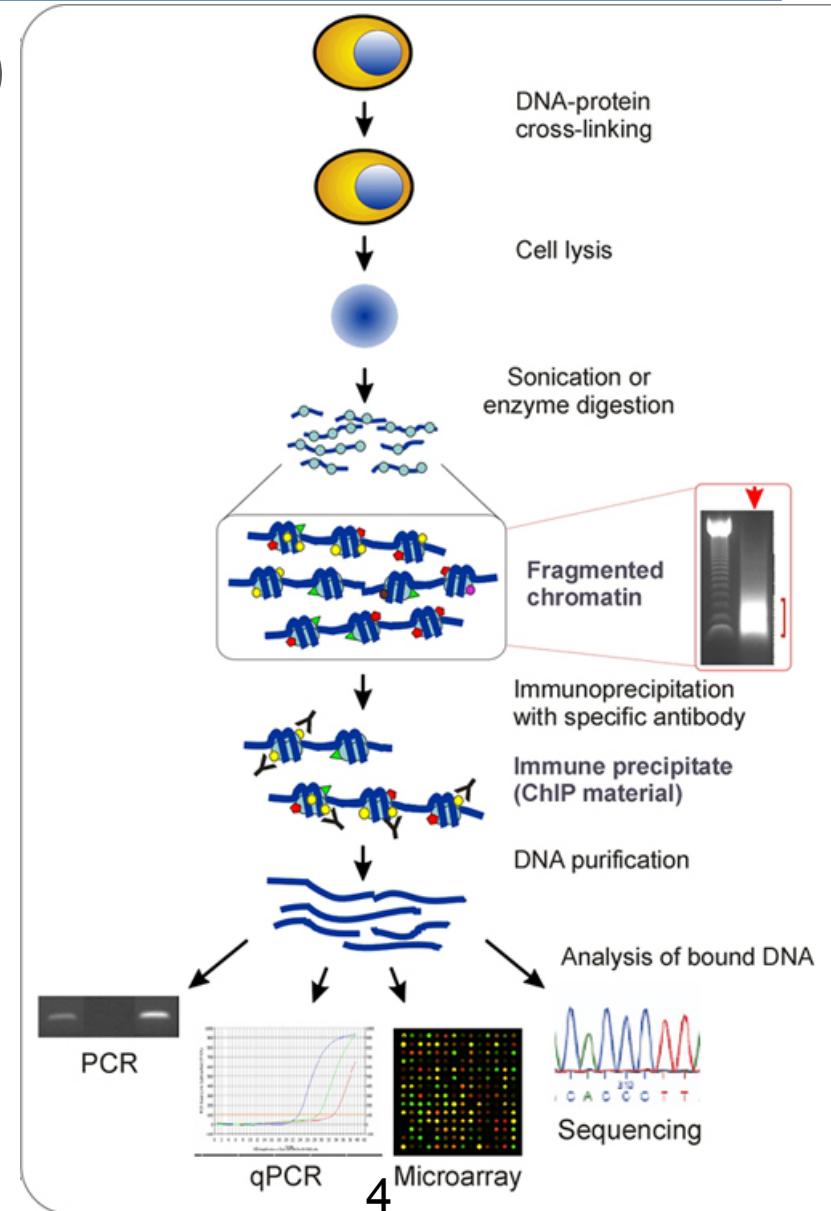
-small-scale: PCR / qPCR

- large-scale:

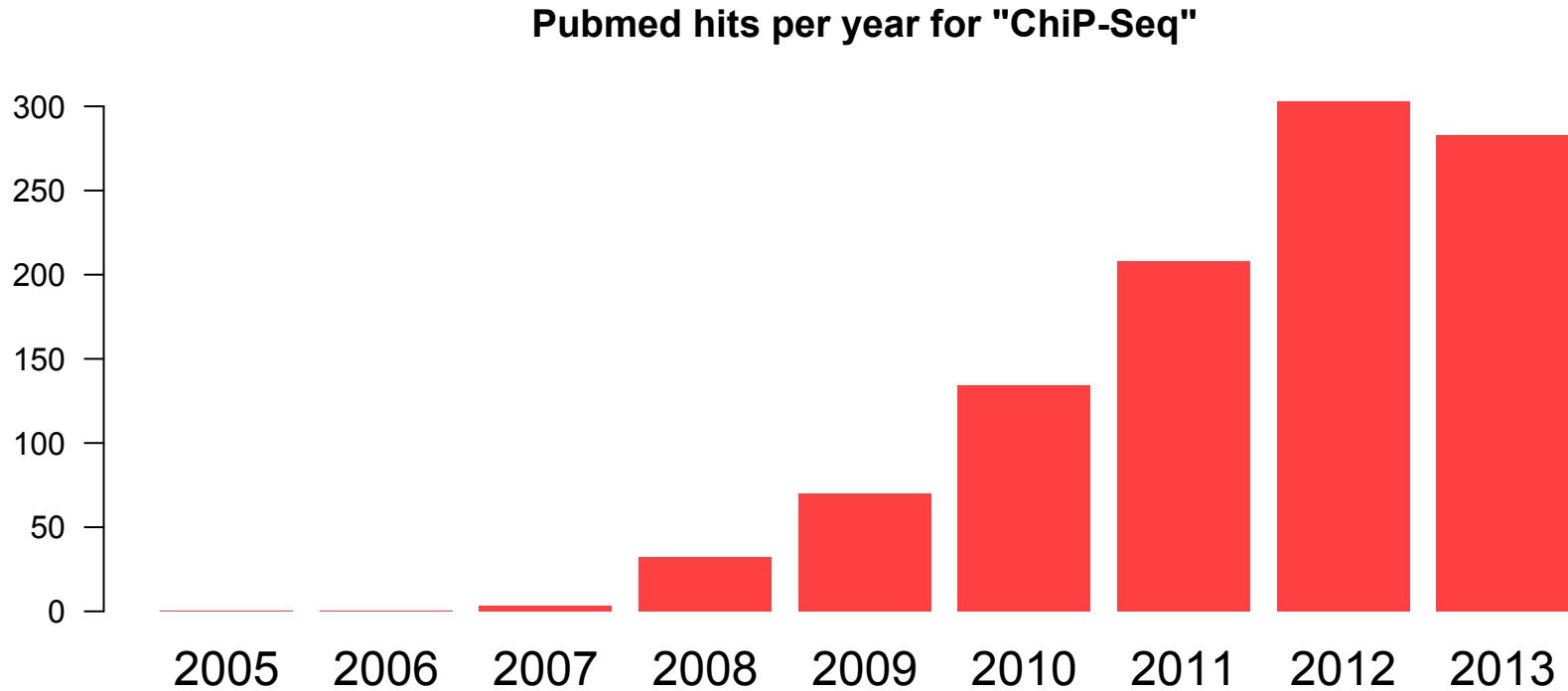
- microarray = ChIP-on-chip
- sequencing = ChIP-seq

Main challenge:

-quality/specificity of the antibodies



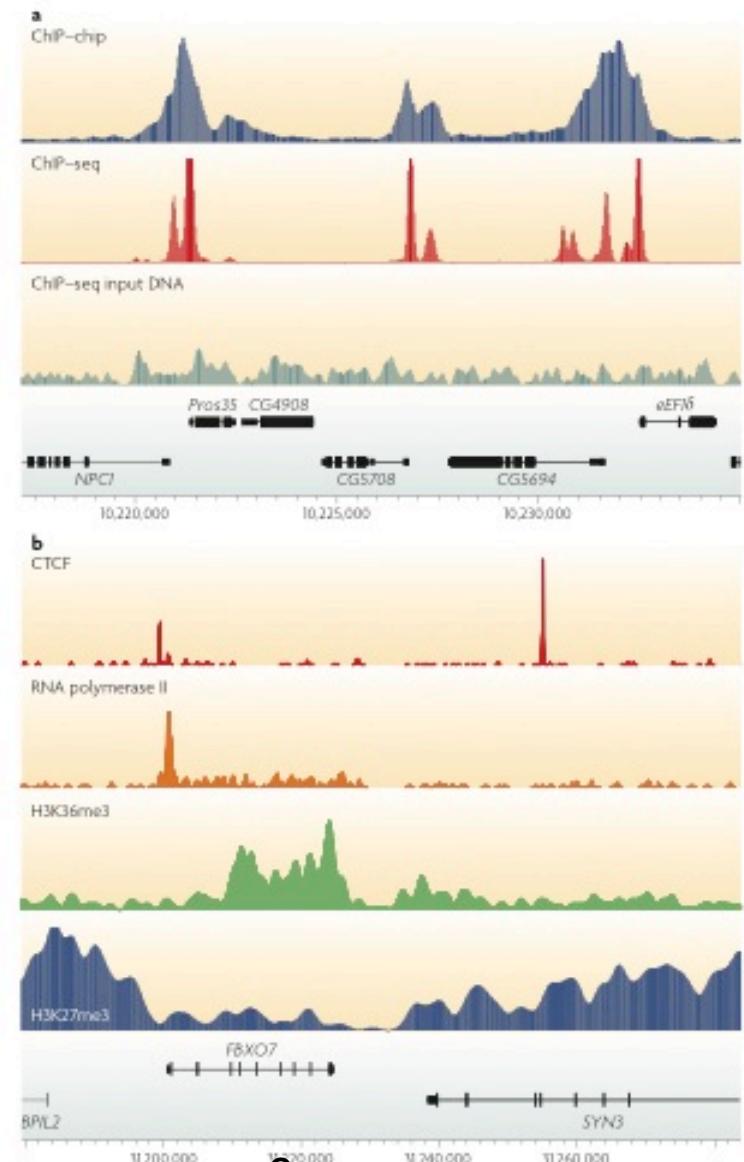
ChIP-seq is a recently-adopted technique !



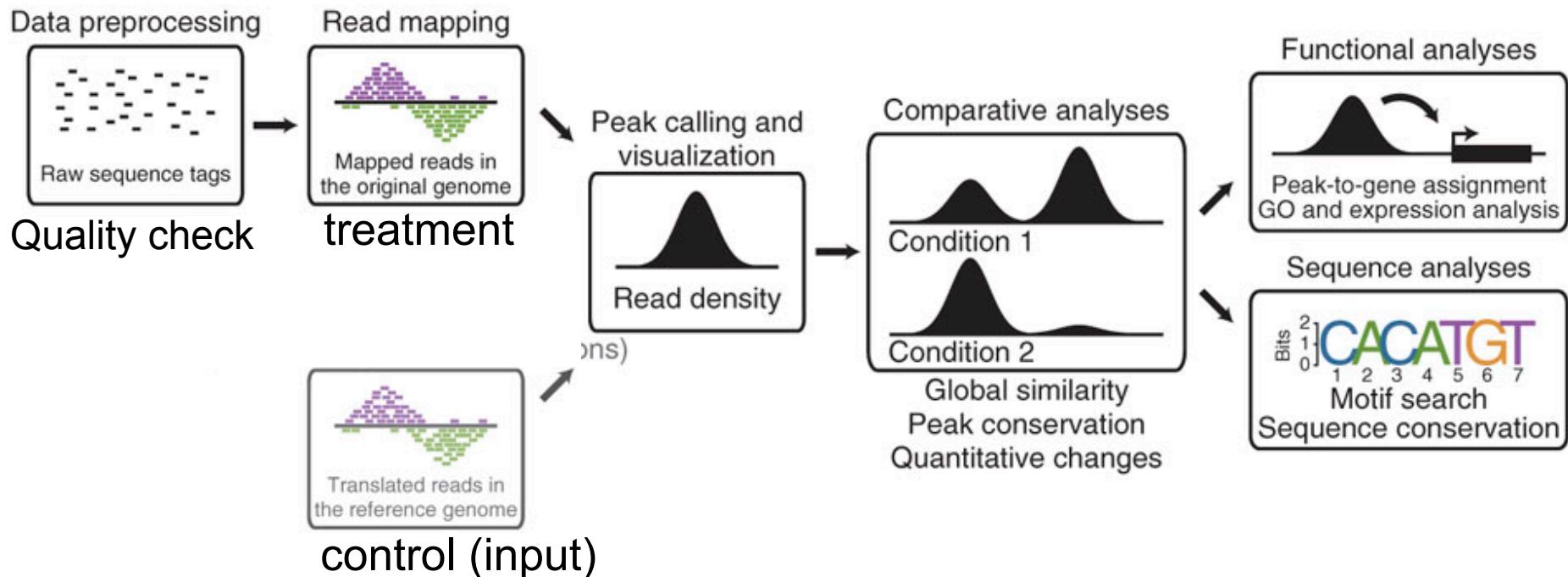
ChIP-seq applications

- find ***all*** regions in the genome bound by
 - a specific **transcription factor**
 - **histones** bearing a specific **modification**
- in a given ***experimental condition*** (cell type, developmental stage,...)

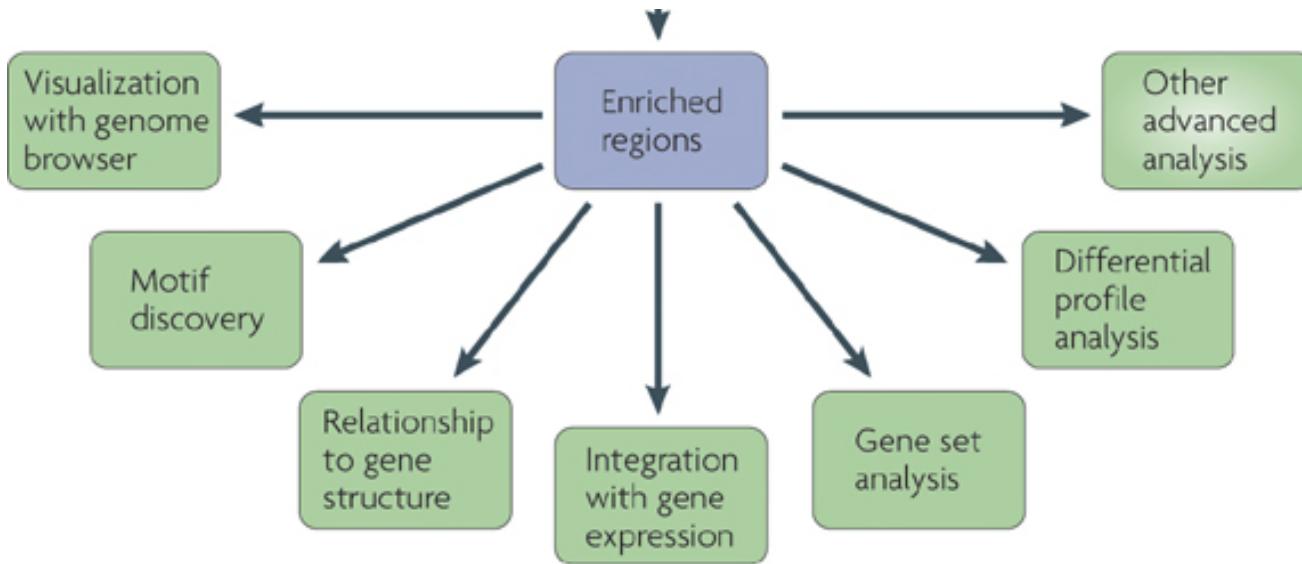
The obtain ChIP-seq **profiles** have **different shapes**, depending on the targeted protein



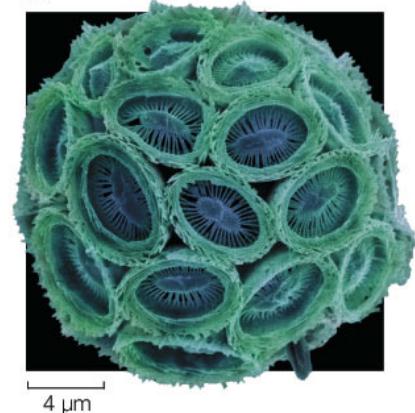
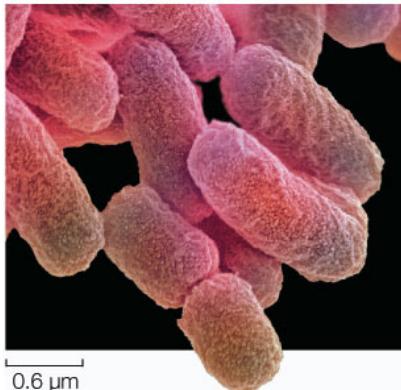
ChIP-seq analysis workflow



ChIP-seq analysis workflow: downstream analyses



Nature Reviews | Genetics



What is the biological question ?



(G) Giant tortoise



Galápagos hawk

What is the biological question ?

« see if you can find something in the data »

What is the biological question ?

« see if you can find something in the data »

What is the biological question ?

- Where do a transcription factor (TF) bind ?
 - ✓ In a specific context (tissue, developmental stage, mutant)
 - ✓ By comparison to another context (WT vs mutant, different time points)

What is the biological question ?

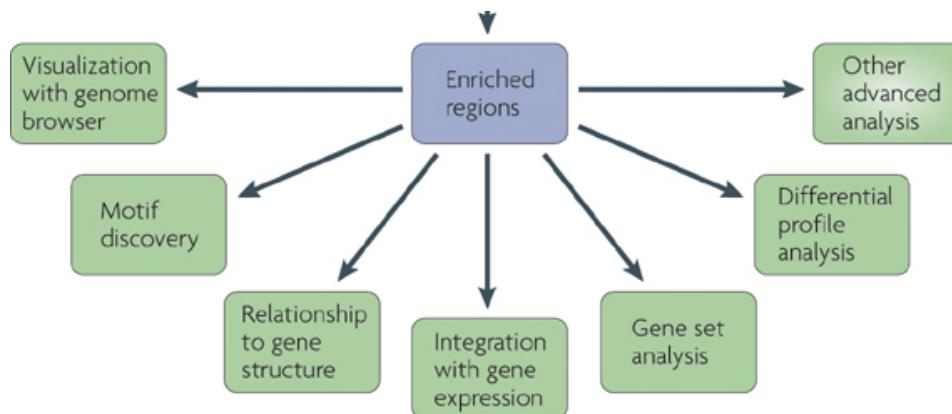
- **Where** do a transcription factor (TF) bind ?
 - ✓ In a **specific context** (tissue, developmental stage, mutant)
 - ✓ By **comparison** to another context (WT vs mutant, different time points)
- **How** do a transcription factor (TF) bind ?
 - ✓ Which **binding motif(s)** (can be several for a given TF !!)
 - ✓ Is the **binding** direct to DNA or via **protein-protein** interactions ?
 - ✓ Are there **cofactors** (maybe affecting the motif !!), and if so, identify them

What is the biological question ?

- Where do a transcription factor (TF) bind ?
 - ✓ In a specific context (tissue, developmental stage, mutant)
 - ✓ By comparison to another context (WT vs mutant, different time points)
- How do a transcription factor (TF) bind ?
 - ✓ Which binding motif(s) (can be several for a given TF !!)
 - ✓ Is the binding direct to DNA or via protein-protein interactions ?
 - ✓ Are there cofactors (maybe affecting the motif !!), and if so, identify them
- Which regulated genes are directly regulated by a given TF ?
- What are the targets of a given TF ?
- Where are the promoters (PolIII) and chromatin marks ?

What is the biological question ?

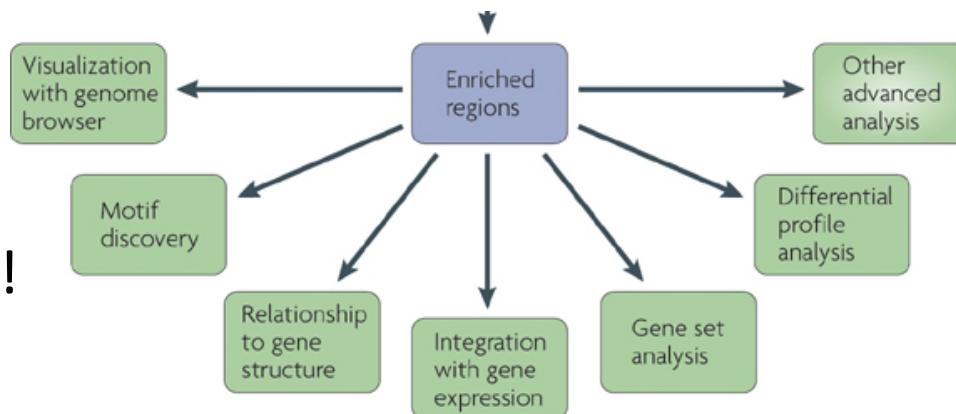
→ Should drive all « downstream » analyses



What is the biological question ?

→ Should drive all « downstream » analyses

Will take time
to « do it all » !!!



What is the biological question ?
What can be the following experimental work ?

What is the biological question ?

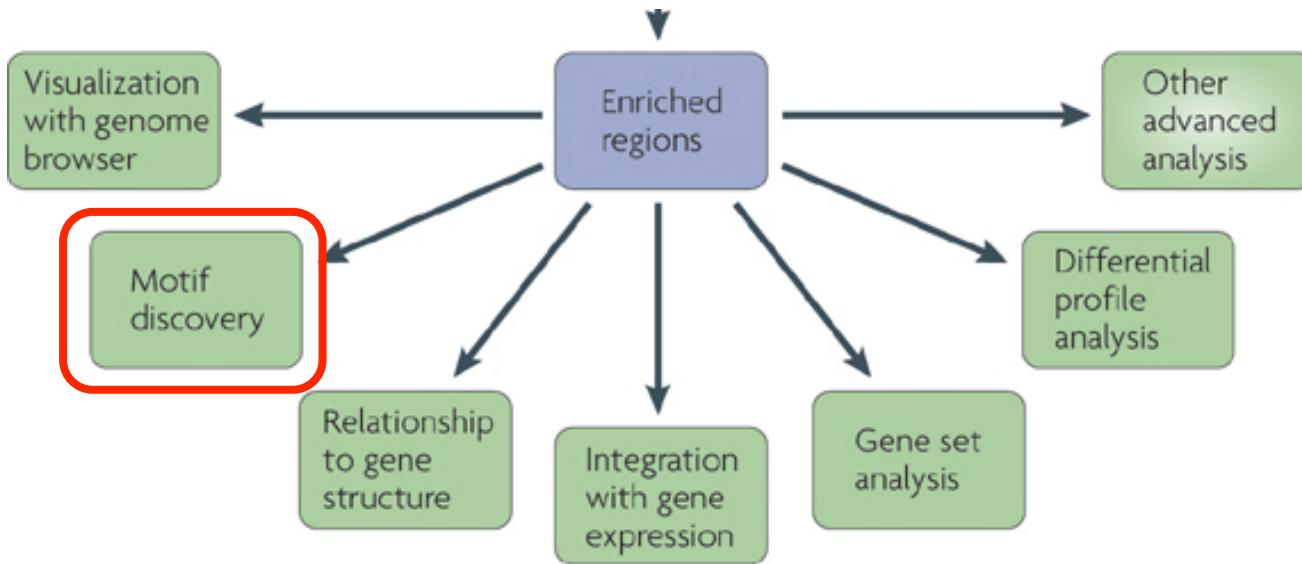
What can be the following experimental work ?

- ➔ cell biology (eg: luciferase assay) ?
- ➔ in vitro assays (eg: EMSA) ?
- ➔ Proteomic (eg: mass spectrometry) ?
- ➔ Transgenics ?
- ➔ Will depend on
 - ✓ the organism
 - ✓ available infrastructure

What is the biological question ?

- Where do a transcription factor (TF) bind ?
 - ✓ In a specific context (tissue, developmental stage, mutant)
 - ✓ By comparison to another context (WT vs mutant, different time points)
- How do a transcription factor (TF) bind ?
 - ✓ Which binding motif(s) (can be several for a given TF !!)
 - ✓ Is the binding direct to DNA or via protein-protein interactions ?
 - ✓ Are there cofactors (maybe affecting the motif !!), and if so, identify them
- Which regulated genes are directly regulated by a given TF ?
- What are the targets of a given TF ?
- Where are the promoters (PolII) and chromatin marks ?

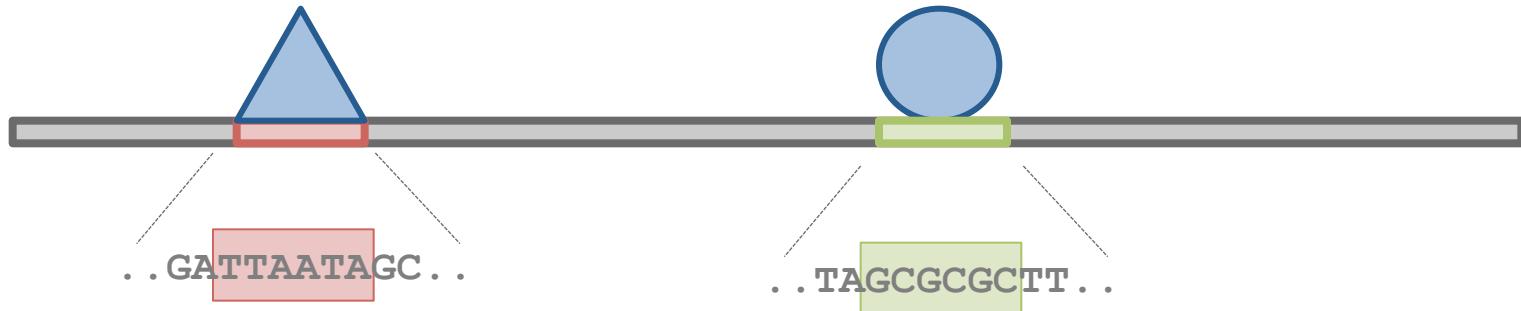
ChIP-seq analysis workflow: downstream analyses



Nature Reviews | Genetics

Transcription factor specificity

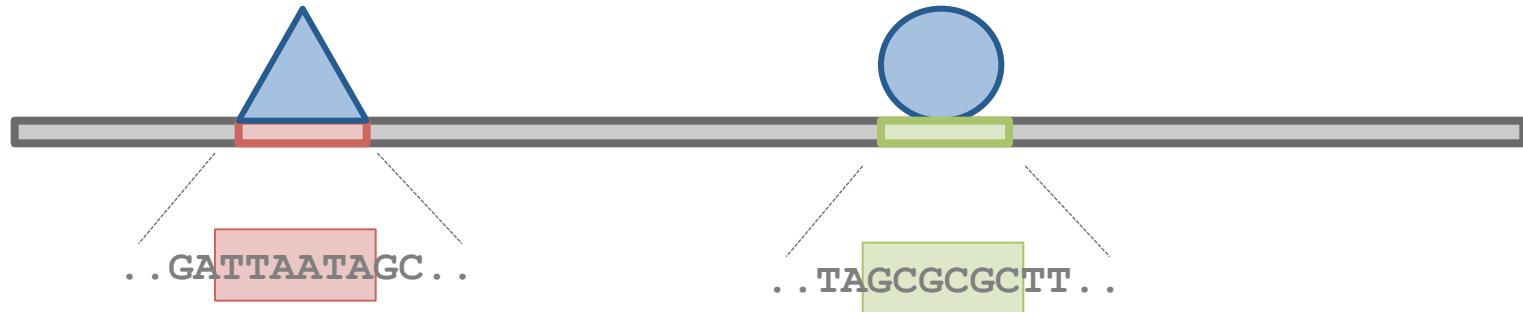
How do TF « know » where to bind DNA ?



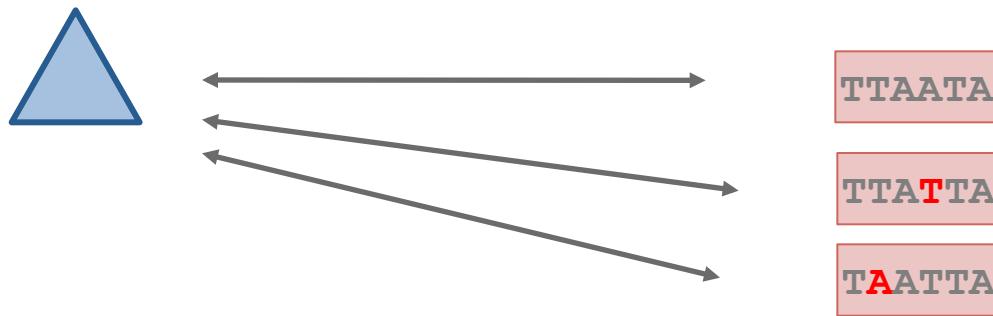
TF recognize TFBS with specific DNA sequences

Transcription factor specificity

How do TF « know » where to bind DNA ?



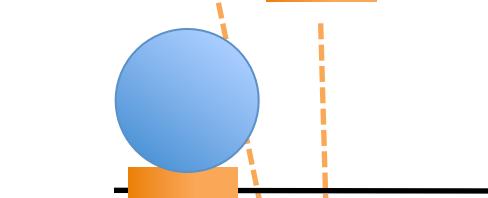
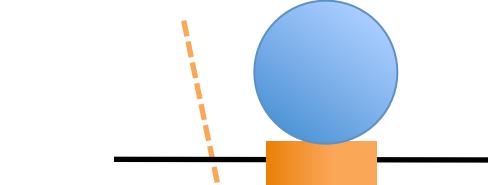
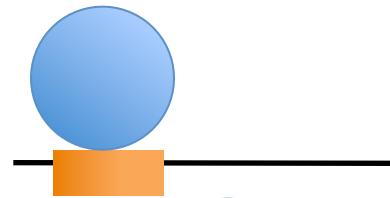
TF recognize TFBS with specific DNA sequences



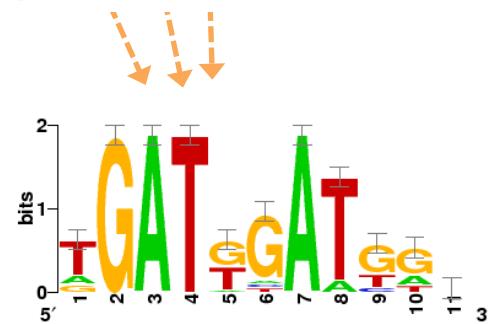
TFBSs are *degenerate*:
a given TF is able to bind DNA on TFBSs with different sequences

Binding specificity of a given TF

transcription factor



cis-regulatory elements

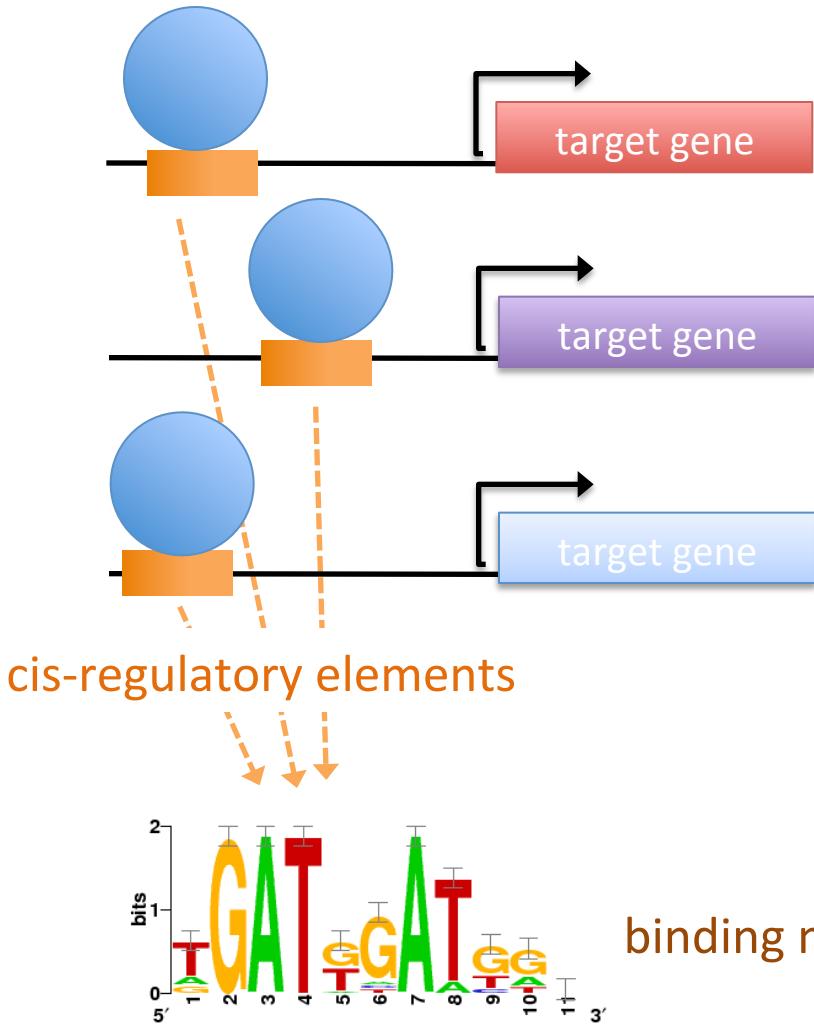


binding motif (represented as a sequence logo)

23

de novo motif discovery

transcription factor



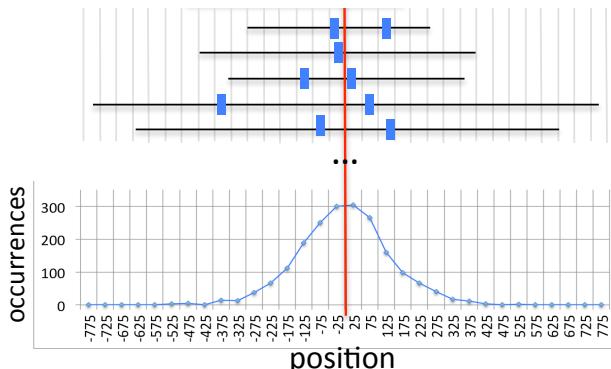
Problem :

*How can we model/describe
the binding specificity of
a given TF ?*

*If there is a common regulating
factor, can we discover its motif
only using these sequences ?*

de novo motif discovery

- Find exceptional motifs based on the sequence only
(*A priori* no knowledge of the motif to look for)
- Criteria of exceptionality:
 - higher/lower frequency than expected by chance
(over-/under-representation)
 - concentration at specific positions relative to some reference coordinate
(positional bias)



de novo motif discovery

- Tools already exist for a long time !

- MEME (1994)
 - RSAT oligo-analysis (1998)
 - AlignACE (2000)
 - Weeder (2001)
 - MotifSampler (2001)

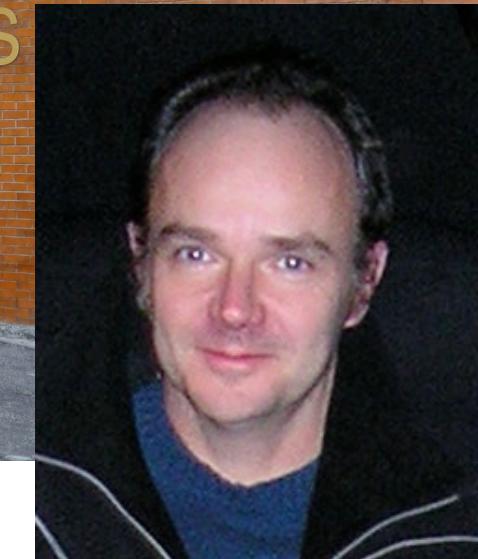
de novo motif discovery

- Tools already exist for a long time !

- MEME (1994)
 - **RSAT oligo-analysis (1998)**
 - AlignACE (2000)
 - Weeder (2001)
 - MotifSampler (2001)

Regulatory Sequence Analysis Tools (RSAT)

- Since 1998 (15 years !)
- Initiated in Cuernavaca, Mexico
- yeast cis-regulatory elements



Jacques van Helden



RSAT developers/contributors over the years



Sylvain Brohée
Postdoc



Nicolas Simonis
Postdoc



Didier Croes
Postdoc



Didier Gonze
Premier assistant



Carl Herrman
MCU



Ariane Toussaint
Professor Emeritus



Jacques van Helden
Professor



Leon Juvenal
Hajingaboe
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

- **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 Hajingaboe*)
- 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*)



Guy Bottu
Postdoc



Lionel Spinelli
Engineer

RSAT improvements over the years

The screenshot shows the RSAT homepage. At the top left is the RSAT logo. To its right is the NeAT logo. Below them is a banner for BiGRE - ULB. The main title "Regulatory Sequence Analysis Tools" is centered above a navigation bar with links: Tool Map, Introduction, Forum, Tutorials, Publications, Credits, People, Data, Download. A sidebar on the left lists "Most popular tools" including "retrieve sequence", "retrieve Ensembl seq", "oligo-analysis (words)", "matrix-scan (quick)", and "random sequence". Other sections in the sidebar include Genomes and genes, Sequence tools, Matrix tools, Build control sets, Motif discovery, Pattern matching, Comparative genomics, NGS - ChIP-seq, Conversion/Utilities, Drawing, and SOAP Web services. A "New programs" section is also present. A central text area welcomes users to RSAT, mentioning the 3rd Tutorial at ECCB 2010 and RSS feed. A "Warnings" box highlights "Vertebrate genomes". At the bottom, there's a section for "Regulatory Sequence Analysis Tools - Web servers" with icons for Brussels (Belgium), Brussels (2) (Belgium), Cuernavaca (Mexico), Uppsala (Sweden), Marseille TAGC (France), and ENS Paris (France), each with a corresponding URL.



www.rsat.eu

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

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

Medina-Rivera, Abreu-Goodger, Thomas-Chollier, Salgado, Collado-Vides, van Helden **Nucleic Acids Research**, 2011

Sand, Thomas-Chollier, van Helden **Bioinformatics**, 2009

Thomas-Chollier*, Sand*, Turatsinze, Janky, Defrance, Vervisch, van Helden **Nucleic Acids Research**, 2008

Sand, Thomas-Chollier, Vervisch, van Helden **Nature Protocols**, 2008

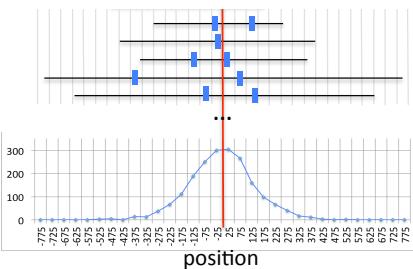
Thomas-Chollier*, Turatsinze*, Defrance, van Helden **Nature Protocols**, 2008

van Helden, **Nucleic Acids Research**, 2003

van Helden, André, Collado-Vides **Yeast**, 2000

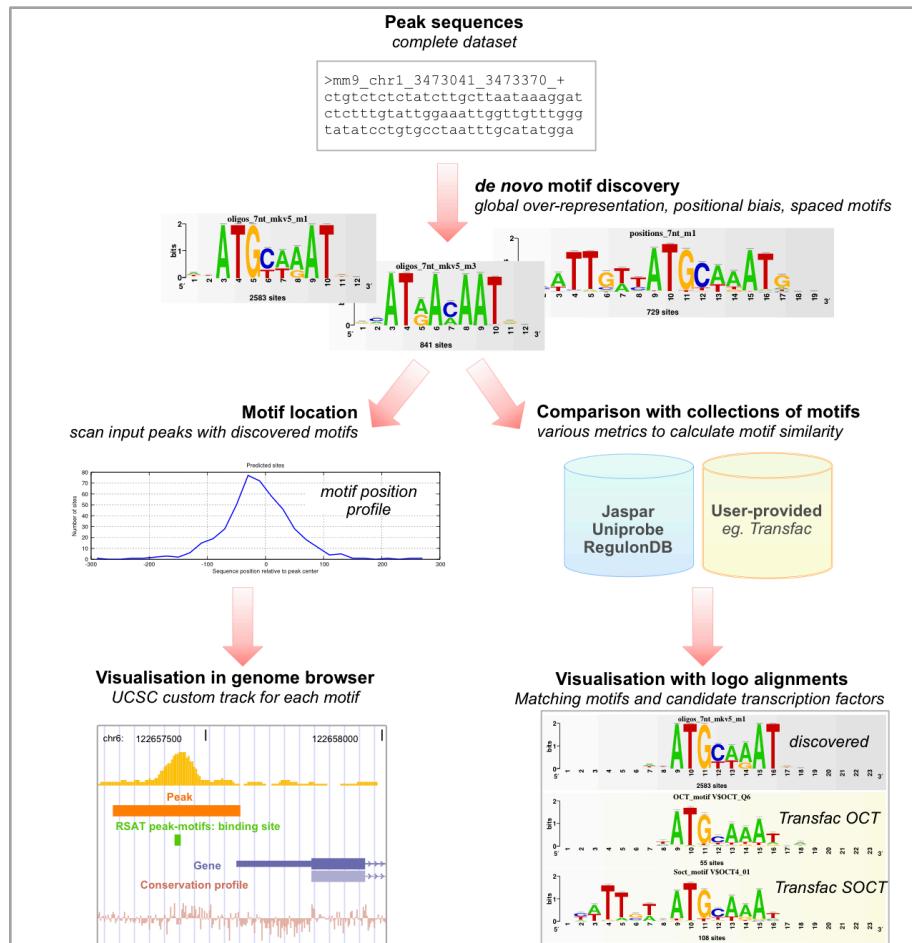
New approaches for ChIP-seq datasets

- **Size, size, size**
 - limited numbers of promoters and enhancers
 - ↓
 - dozens of thousands of peaks !!!!!!
- **the problem is slightly different**
 - promoters: 200-2000bp from co-regulated genes
 - ↓
 - peaks: 300bp, positional bias
- **motif analysis: not just for specialists anymore !**
 - complete user-friendly workflows



New approaches for ChIP-seq datasets

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



Thomas-Chollier et al Nucleic Acids Research, 2012

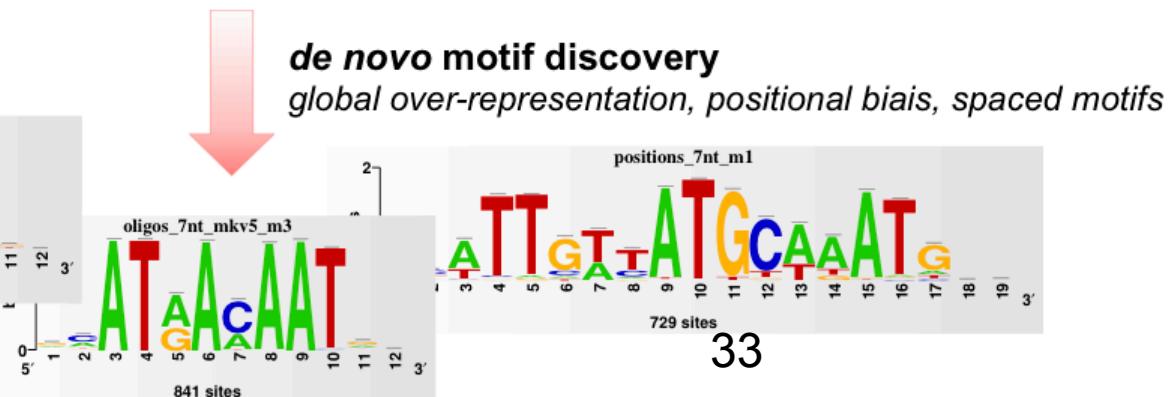
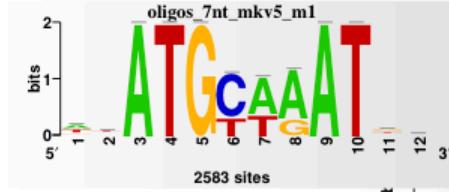
Peaks coordinates



chr1	3002142	3002195
chr1	3002804	3002853

Peak sequences complete dataset

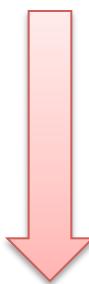
```
>mm9_chr1_3473041_3473370_+  
ctgtctcttatcttgcttaataaaggat  
ctctttgtattggaaattgggttgttggg  
tatatcctgtgcctaatttgcataatgga
```



Peaks coordinates



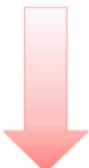
chr1	3002142	3002195
chr1	3002804	3002853



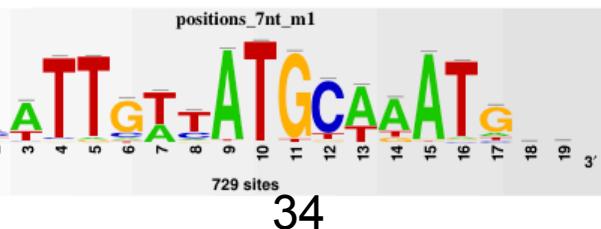
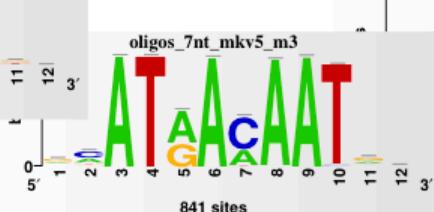
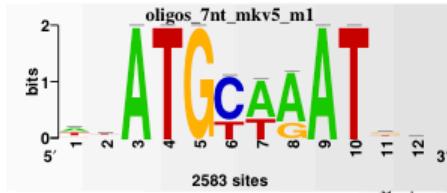
Extract corresponding sequences

Peak sequences complete dataset

```
>mm9_chr1_3473041_3473370_+  
ctgtctcttatcttgcttaataaaggat  
ctctttgtattggaaattgggtgttggg  
tatatcctgtgcctaatttgcataatgga
```

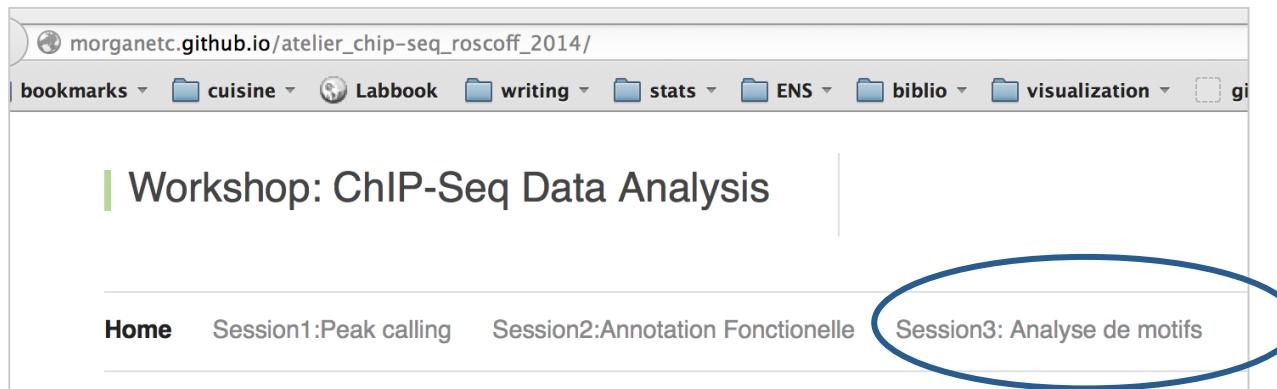


de novo motif discovery
global over-representation, positional bias, spaced motifs



Hands on !

- Go to the companion website
 - <http://ecole-bioinfo-aviesan.sb-roscoff.fr/>
 - Section “chip-seq”



- Follow all steps of **Retrieving sequences from your peaks**

Using RSAT

The screenshot shows the RSAT (Regulatory Sequence Analysis Tools) web interface. At the top left is the RSAT logo and a NeAT button. Below is a sidebar with sections like 'Most popular tools' (with options like 'retrieve sequence', 'oligo-analysis (words)', etc.), 'Genomes and genes' (marked 'New!'), 'Sequence retrieval' (marked 'New!'), 'Pattern discovery', 'Pattern matching', 'Comparative genomics', 'Conversion/Utilities', 'Drawing', 'Web services' (all marked 'New!'), 'Help' (with 'Map of the tools', 'Introduction', 'Tutorials', 'Course', 'Contact & Forum' marked 'New!'), and 'Information' (with a feedback link). The main content area shows an 'Expandable menus' section with a large arrow pointing to a detailed 'Sequence retrieval' menu on the right, which includes options like 'retrieve sequence', 'retrieve EnsEMBL sequence' (marked 'New!'), 'purge sequence', 'convert sequence', and 'random sequences'. Below this, three grey arrows point from the sidebar sections to numbered steps: '2. Run the analysis', '3. Visualization', and 'Help: tutorials, forum'. Two orange arrows point from the 'Help' and 'Information' sections to 'Help: tutorials, forum' and 'Information: publications,...' respectively.

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
New!

► 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

RSA-tools - retrieve sequence ← **Tool name**

Returns upstream, downstream or ORF sequences for a list of genes ← **Tool description**

Remark: If you want to retrieve sequences from an organism that is in the Ensembl database, we recommend to use the [retrieve-ensembl-seq](#) program instead

Single organism Organism: Saccharomyces cerevisiae ↴
 Multiple organisms

Genes: all selection

Upload gene list from file [Browse...](#) ← **Tool parameters**
 Query contains only IDs (no synonyms)

Feature type: CDS mRNA tRNA rRNA scRNA
Sequence type: upstream From: default To: default

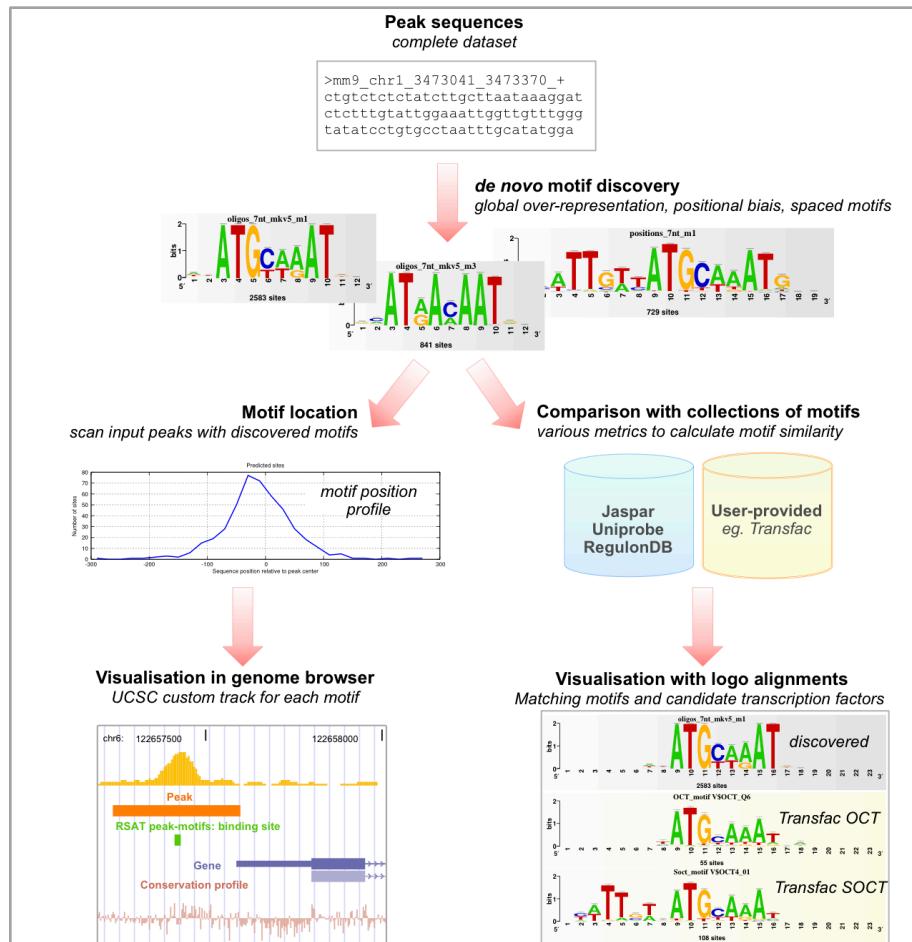
 Prevent overlap with neighbour genes (noorf)
 Mask repeats (only valid for organisms with annotated repeats)
 Admit imprecise positions
Sequence format: fasta
Sequence label: gene name
Output: server display email

Output
Go button (launches the analysis)
Demo button (fill in the form for test purposes)
Help

GO DEMO MANUAL TUTORIAL MAIL

New approaches for ChIP-seq datasets

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



Thomas-Chollier et al *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	-
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

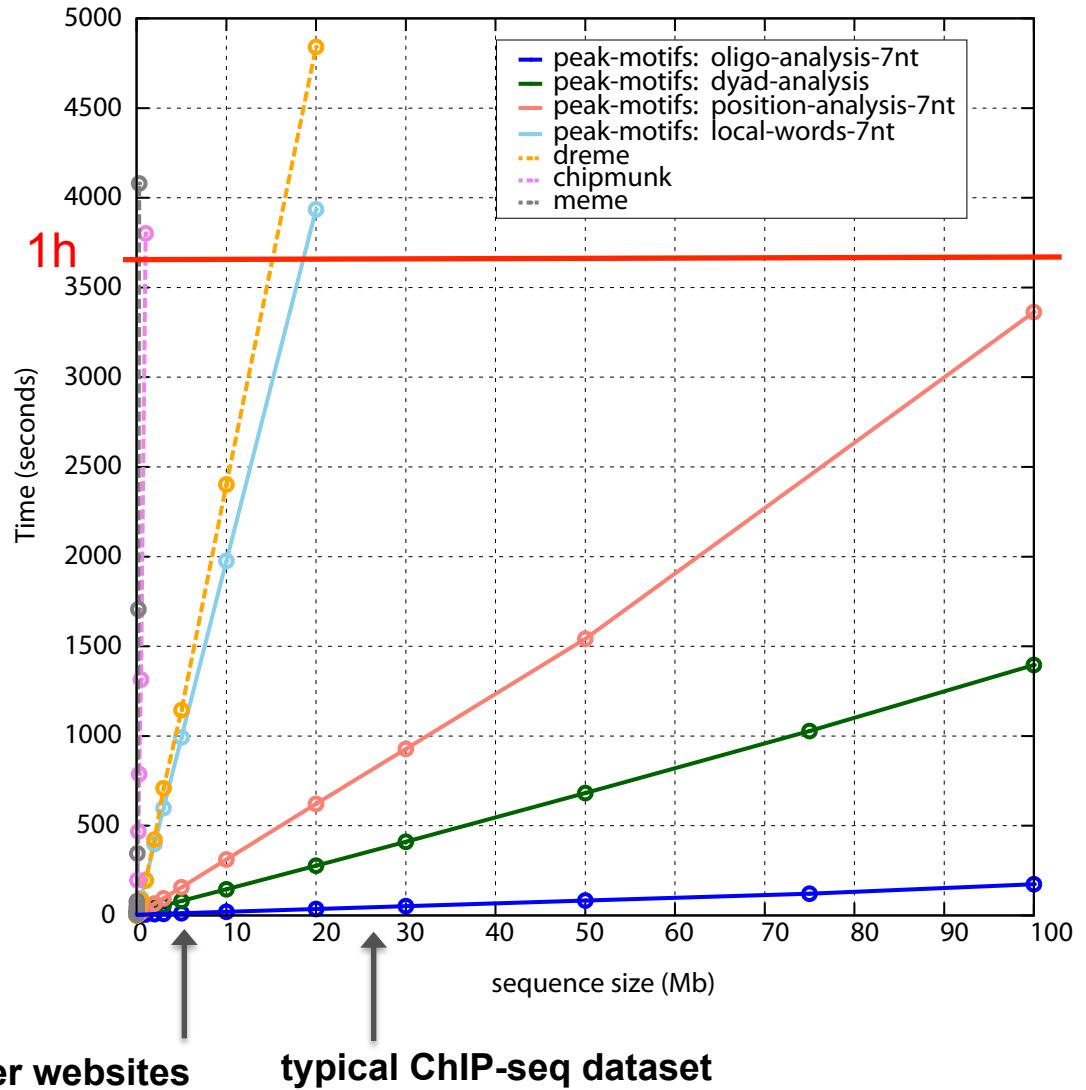
Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	100kb (web site)	500kb (web site)	unrestricted, but motif discovery restricted to 600 peaks clipped to 100bp	motif discovery restricted to a few hundred base pairs	-
peak finding	no	no	no	yes	no	
annotation of peak-flanking genes	no	no	yes	no	no	
sequence composition (mono- and di-nucleotides)	yes	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	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

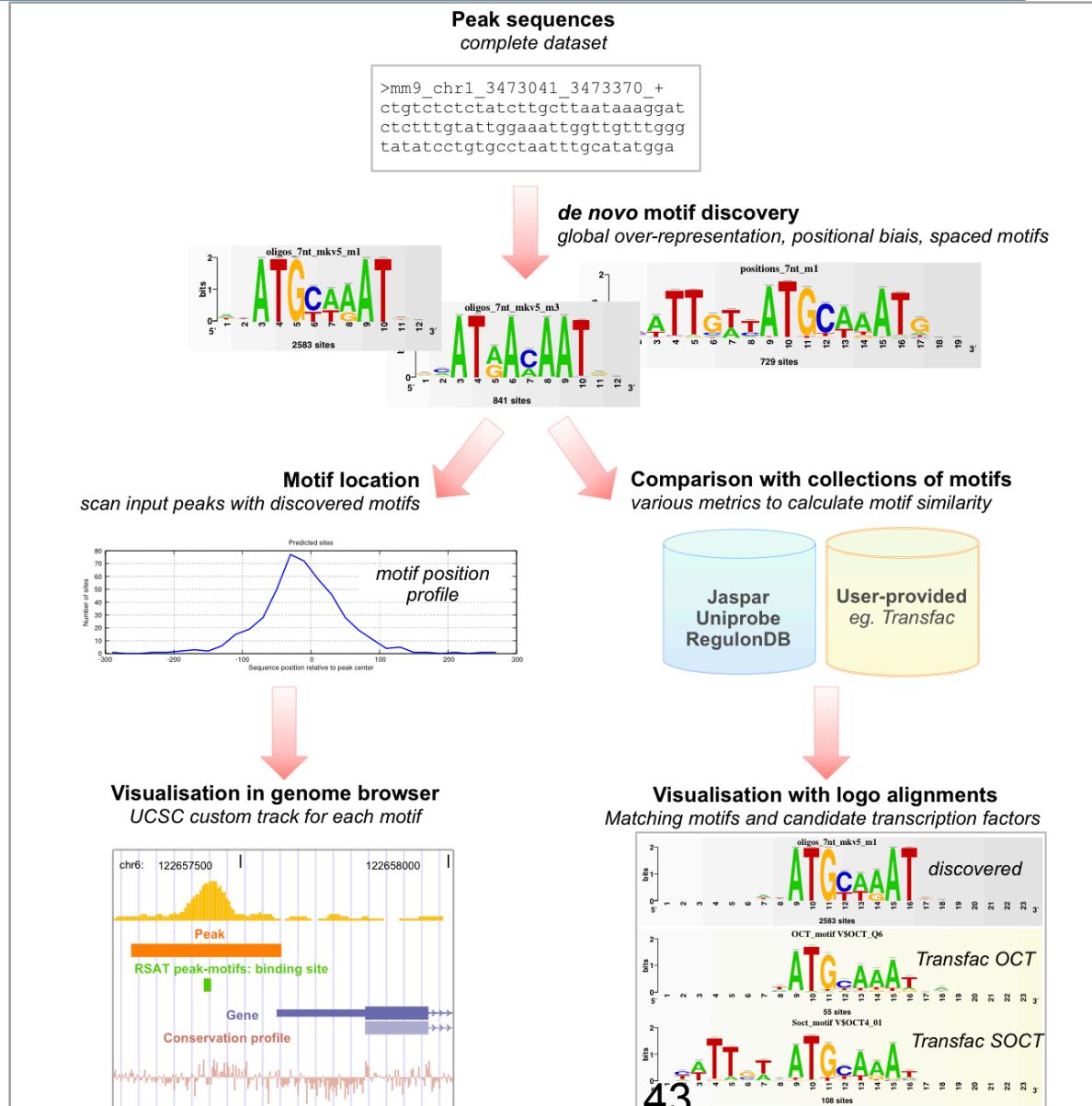
Peak-motifs: why providing yet another tool ?

- fast and scalable
- treat full-size datasets



Peak-motifs: why providing yet another tool ?

- fast and scalable
 - treat full-size datasets
 - complete pipeline



Peak-motifs: why providing yet another tool ?

- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface

RSA-tools - peak-motifs

Pipeline for discovering motifs in massive ChIP-seq peak sequences.

Conception², implementation¹ and testing³: Jacques van Helden^{cit}, Morgane Thomas-Chollier^{cit}, Matthieu Defrance^{cit}, Olivier Sand^b, Denis Thieffry^{cit}, and Carl Herrmann^{cit}.

► Information on the methods used in peak-motifs

Peak Sequences

Title: Kr.D.mel 1-3h Markov m=k-2

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

Or select a file to upload (.gz compressed files supported)
/Kr.D.mel_E01-03h_Eisen_rep1.fasta

Mask: lower

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

Control sequences

Optional: control dataset for differential analysis (test vs control)

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

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

Mask: none

► Reduce peak sequences

► Motif discovery parameters

► Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs

► Locate motifs and export predicted sites as custom UCSC tracks

Output: display email

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

[\[MANUAL\]](#) [\[TUTORIAL\]](#) [\[ASK A QUESTION\]](#)

- accessible to non-specialists

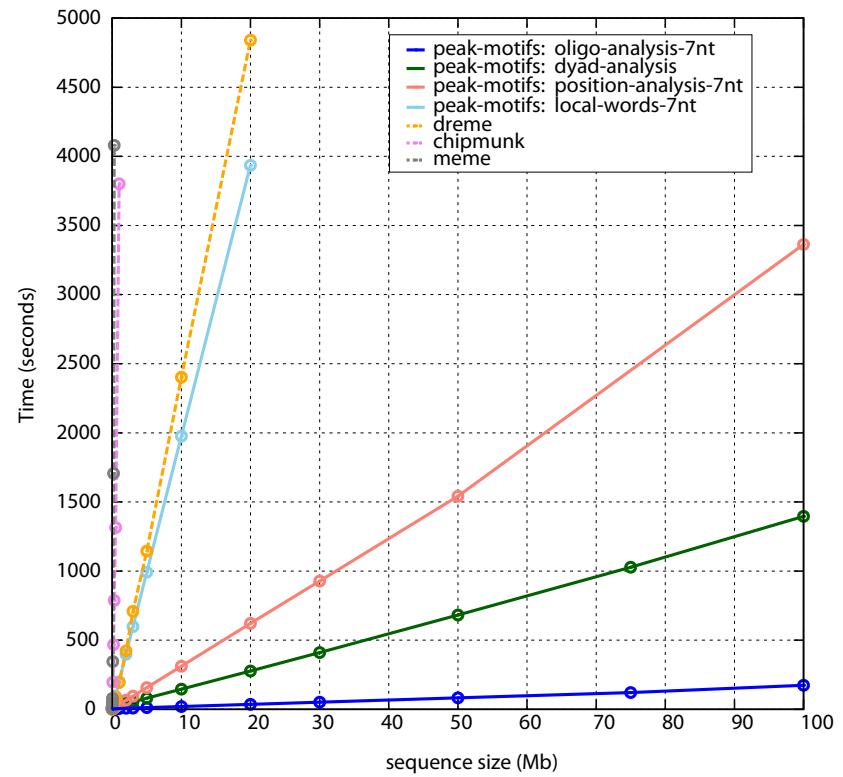
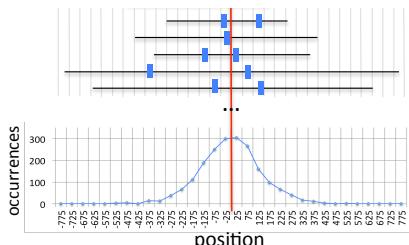
- Demo buttons
- Tutorials & Protocols

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

- HTML report

Peak-motifs: why providing yet another tool ?

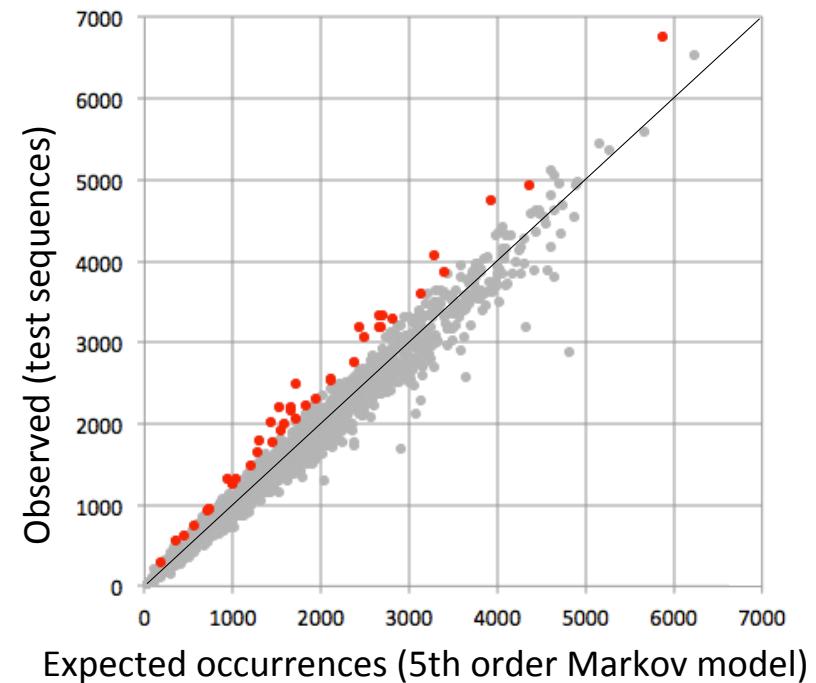
- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface
- accessible to non-specialists
- using 4 complementary algorithms
 - Global over-representation
 - oligo-analysis
 - dyad-analysis (spaced motifs)
 - Positional bias
 - position-analysis
 - local-words



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)	

Observed vs expected 6-mer occurrences

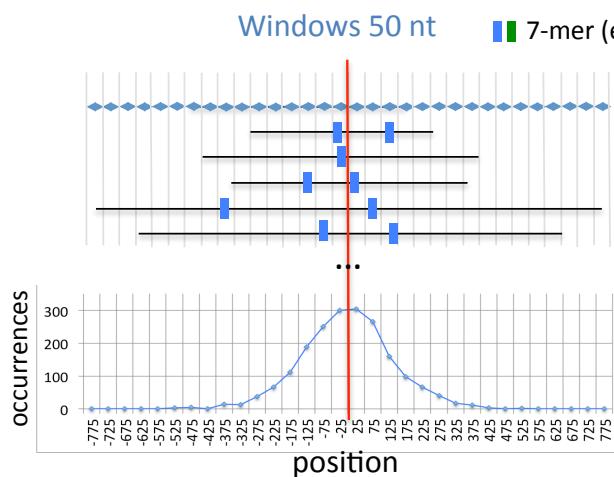


oligo-analysis
dyad-analysis (spaced motifs)

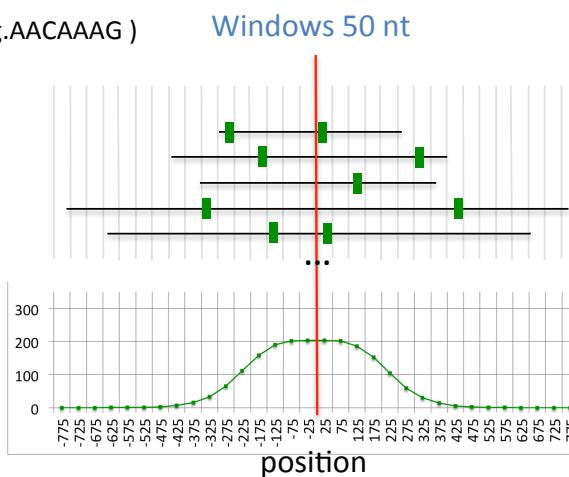
Motif discovery methods: positional bias

B

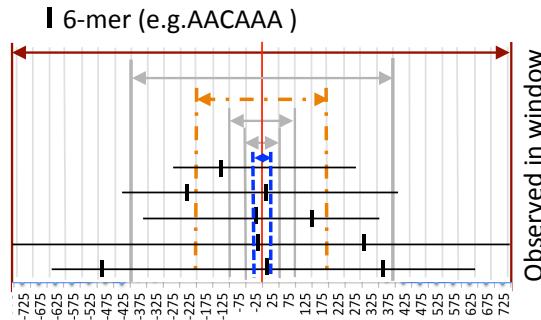
Observed occurrences per window



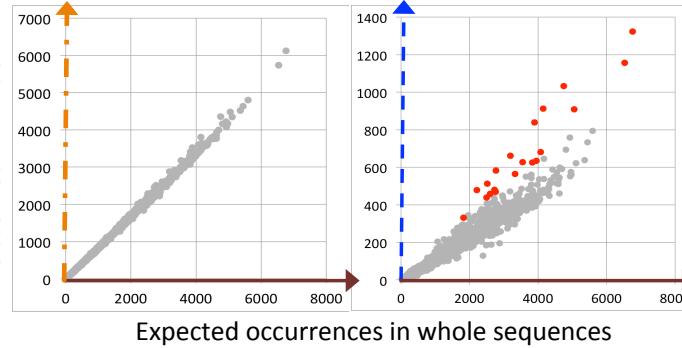
Expected occurrences per window following an homogeneous model



C Observed occurrences per window



Observed vs expected 6-mer occurrences

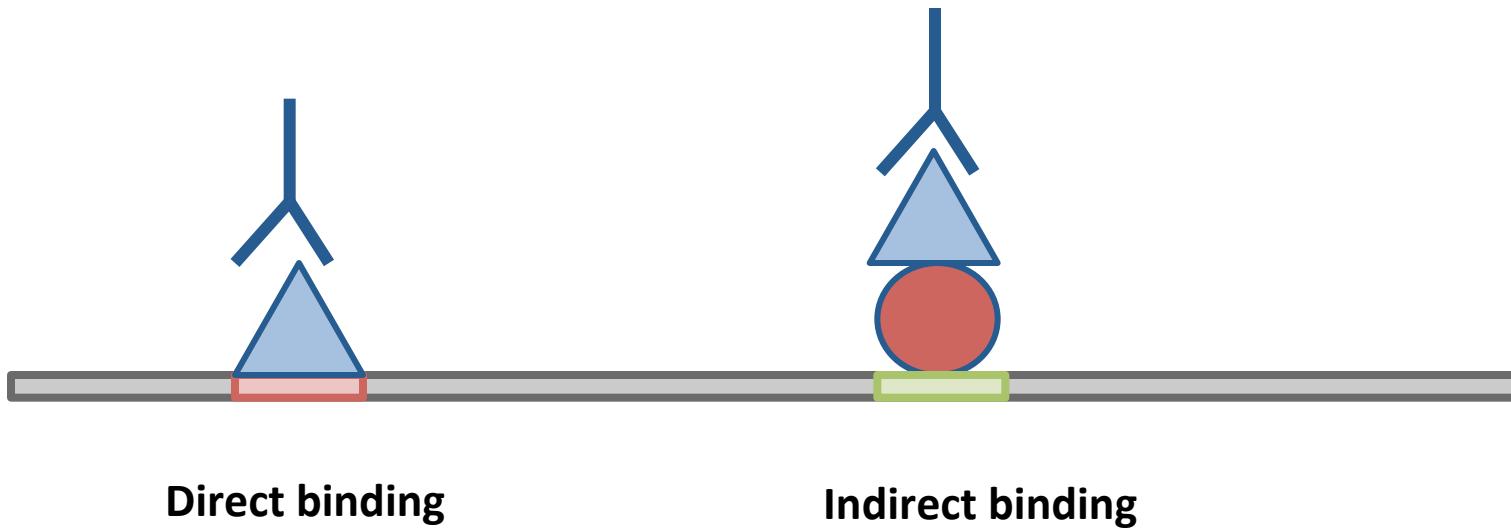


position-analysis

local-words

Direct versus indirect binding

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

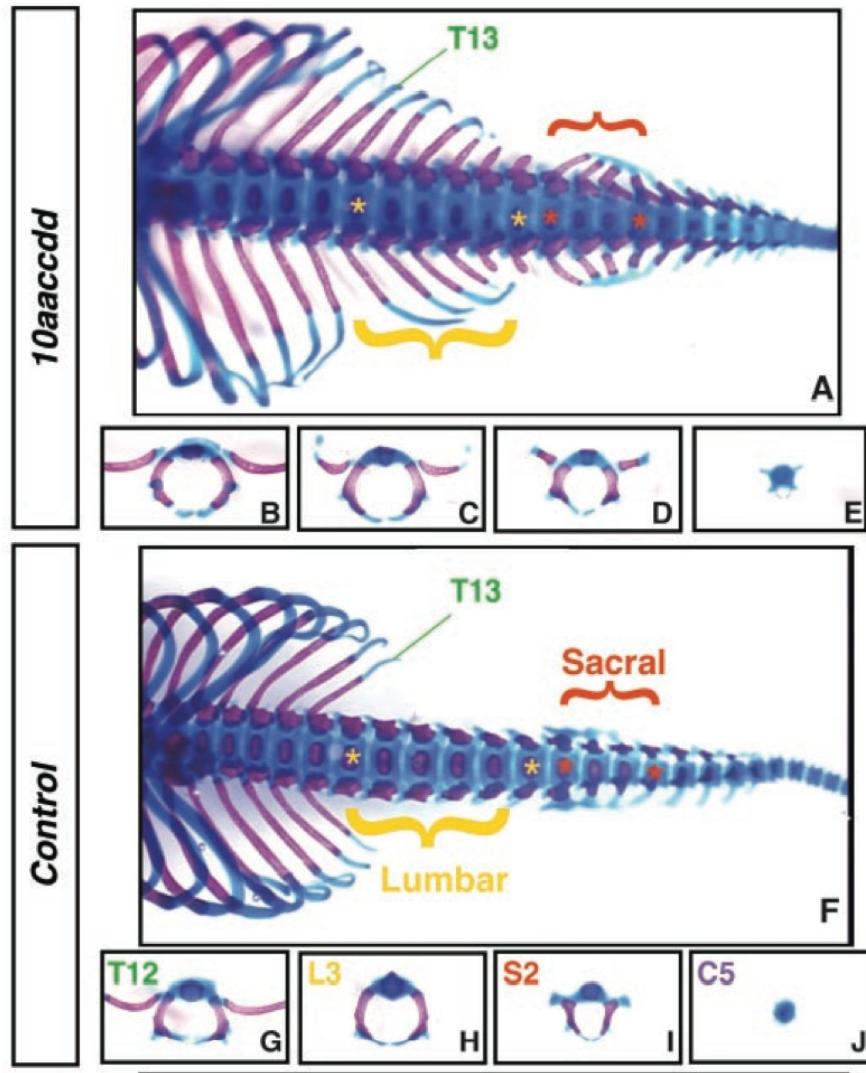


- The motif of the targeted TF is not always found in peaks !

Hands on !

- Go to the companion website
- Follow **steps 1 & 2 of Discovering motifs from peak sequences**

Negative Controls in biology



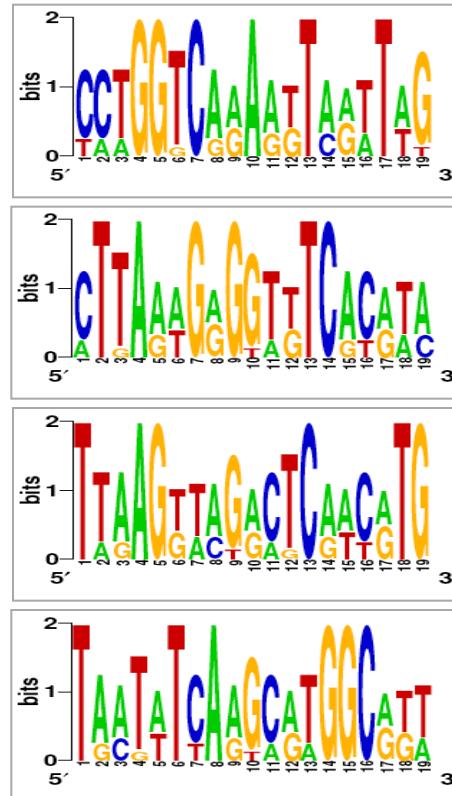
Wellik and Mario R Capecchi, Science, 2003

In the context of cis-regulation

Use different set of *sequences*

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAGAAAAGAGTCAGACATCGAACATACAT ...*HIS7*
5' - ATGGCAGAACATCACTTAAACGTGGCCCCACCCGCTGCACCCGTGACATTGACTCGAAATGACTCAACG ...*ARO4*
5' - CACATCCAACGAATCACCTCACCGTTATCGTACTCACTTCCTCGCATGCCGAAGTGCATAAAAATATTTTT ...*ILV6*
5' - TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATTTCGACAAAATGTATAGTCATTCTATC ...*THR4*
5' - ACAAAAGGTACCTCCTGCCAATCTCACAGATTAATATAGTAAATTGTATGCATATGACTCATCCCACATGAAA ...*ARO1*
5' - ATTGATTGACTCATTTCCTCTGACTACTACCAGTCAAAATGTTAGAGAAAATAGAAAAGCAGAAAAATAAATAA ...*HOM2*
5' - GGCGCCACAGTCCCGTTGGTTATCCGGCTGACTCATTCTGACTCTTTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

Use different set of *matrices*



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 ?

The screenshot shows the header of the Nature journal website. The word "nature" is in large white letters, with "International weekly journal of science" in smaller text below it. A navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. Below this is a breadcrumb trail: Archive > Volume 513 > Issue 7518 > Retractions > Article. The main title of the article is "Retraction: Genomic organization of human transcription initiation complexes".

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

> view all tools

- ▶ Genomes and genes
- ▶ Sequence tools
- ▶ Matrix tools !
- ▼ Build control sets !
 - random gene selection
 - random sequence
 - random genome fragments !
 - random-motif !
 - permute-matrix !
 - random-sites !
 - implant-sites !

Hands on !

- Go to the companion website
- Follow **steps 3 of Discovering motifs from peak sequences**

Hands on !

- Go to the companion website
- Follow all steps 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



5' - TCTCTCTCACGGCTAATTAGGTGATCATGAAAAAAATGAAAAATTCA	TGAG AAAAGAGTC A	GACATCGAACATACAT	... <i>HIS7</i>
5' - ATGGCAGAACATCACTTAAAACGTGGCCCACCCGCTGCACCCTGTG	CATTTGTACGTTACTGCG AAATGACTCA ACG	...	<i>ARO4</i>
5' - CACATCCAACGAATCACCTCACCGTTATCG TGACTCACTT TCTTTCGCATGCCGAAGTGCATAAAAAATATTTTT	<i>ILV6</i>
5' - TGCGAAC AAAAGAGTC ATTACAACGAGGAAATAGAAGAAAATGAAAAATTTCGACAAAATGTATAGTCATTCTATC	<i>THR4</i>
5' - ACAAAAGGTACCTTCCTGGCAATCTCACAGATTAAATAGTAAATTGTCATGCATA TGACTCATCC CGAACATGAAA	<i>ARO1</i>
5' - ATTGAT TGACTCATTT CCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAATAGAAAAGCAGAAAAATAAATAA	<i>HOM2</i>
5' - GGCGCCACAGTCCCGTTGGTTATCCGGC TGACTCATTCTGACTCTTT TTGGAAAGTGTGGCATGTGCTTCACACA	<i>PRO3</i>

- 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:

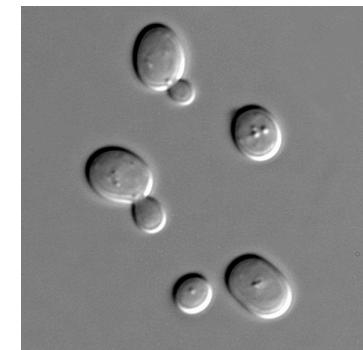
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

■ 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		
aaaaaaa	ttttttt	51
aaaaaag	ctttttt	15
aagaaaa	tttctt	14
gaaaaaa	tttttc	13
tgc当地	ttggca	12
aaaaat	attttt	12
aaatta	taattt	12
agaaaaa	ttttct	11
caagaa	ttcttg	11
aaacgt	acgttt	11
aaagaa	ttcttt	11
acgtgc gcacgt	10	
aataat	attatt	10
aagaag	cttctt	10
atataa	ttatat	10

MET		
aaaaaaa	ttttttt	105
atatat	atatat	41
gaaaaaa	tttttc	40
tatata	tatata	40
aaaaat	attttt	35
aagaaa	tttctt	29
agaaaaa	ttttct	28
aaaata	tatttt	26
aaaaag	cttttt	25
agaaat	atttct	24
aaataa	ttattt	22
taaaaaa	ttttta	21
tgaaaaa	ttttca	21
ataata	tattat	20
atataa	ttatata	20

NIT		
aaaaaaa	ttttttt	80
cttatc gataag	26	
tatata	tatata	22
ataaga	tcttat	20
aagaaa	tttctt	20
gaaaaaa	tttttc	19
atatat	atatat	19
agataa	ttatct	17
agaaaaa	ttttct	17
aaagaa	ttcttt	16
aaaaca	tgtttt	16
aaaaag	cttttt	15
agaaga	tcttct	14
tgataaa	ttatca	14
atataa	tttatat	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	
gaaaaaa tttttc	13	
tgc当地 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	
gaaaaaa tttttc	40	
tatata tatata	40	
aaaaat attttt	35	
aagaaa tttctt	29	
agaaaa ttttct	28	
aaaata tatttt	26	
aaaaag cttttt	25	
agaaaat atttct	24	
aaataa ttattt	22	
taaaaa ttttaa	21	
tgaaaa ttttca	21	
ataata tattat	20	
atataa ttatat	20	

NIT		
aaaaaa tttttt	80	
cttatac gataag	26	
tatata tatata	22	
ataaga tcttat	20	
aagaaa tttctt	20	
gaaaaaa tttttc	19	
atatat atatat	19	
agataa ttatct	17	
agaaaa ttttct	17	
aaagaa ttcttt	16	
aaaaca tgcccc	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:

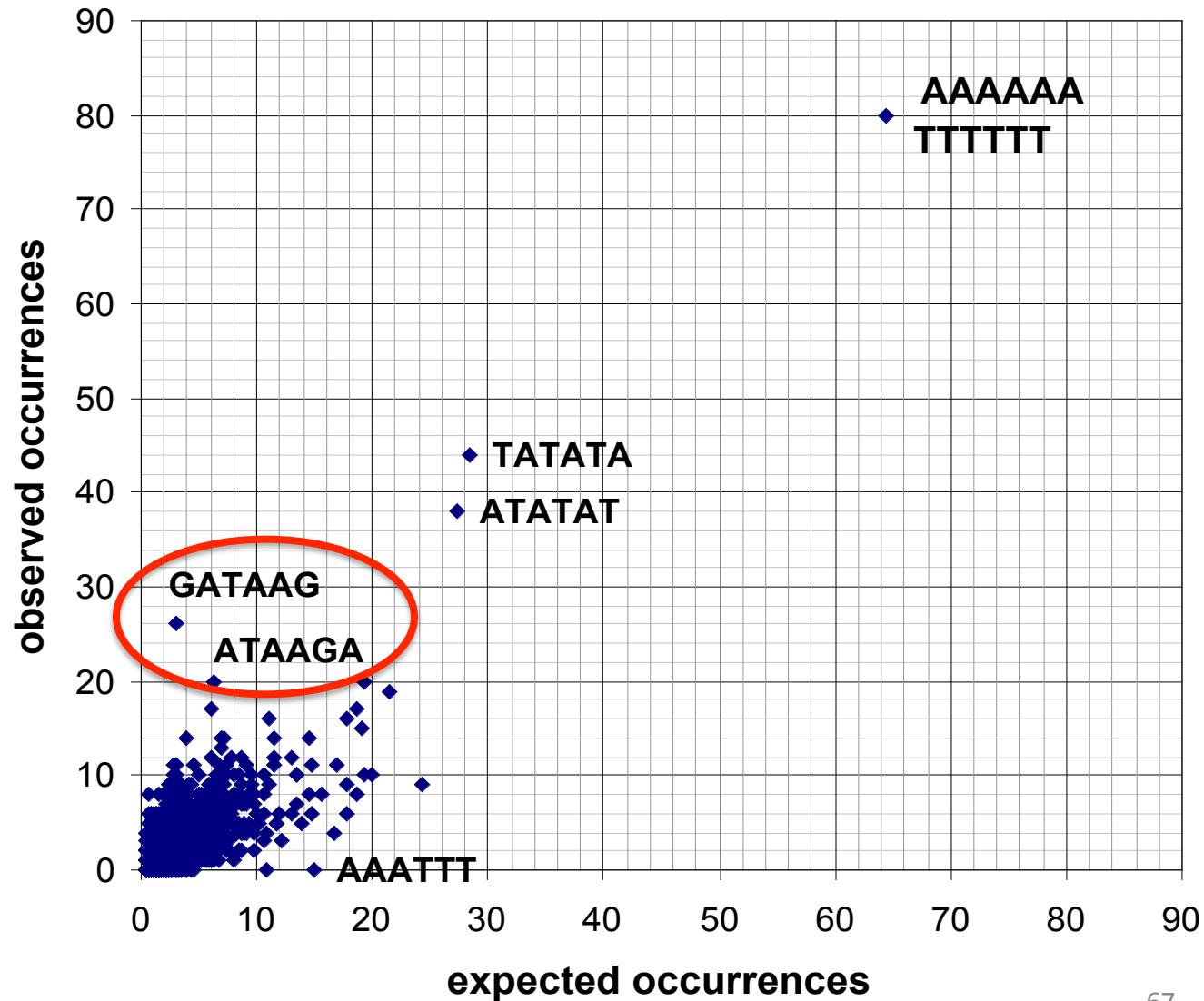
6nt frequencies in the whole set of 6000 yeast **upstream** sequences

;seq	identifier	observed_freq	occ
aaaaaaa	aaaaaa ttttt	0,00510699	14555
aaaaaac	aaaaac gtttt	0,00207402	5911
aaaaaag	aaaaag ctttt	0,00375191	10693
aaaaaat	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

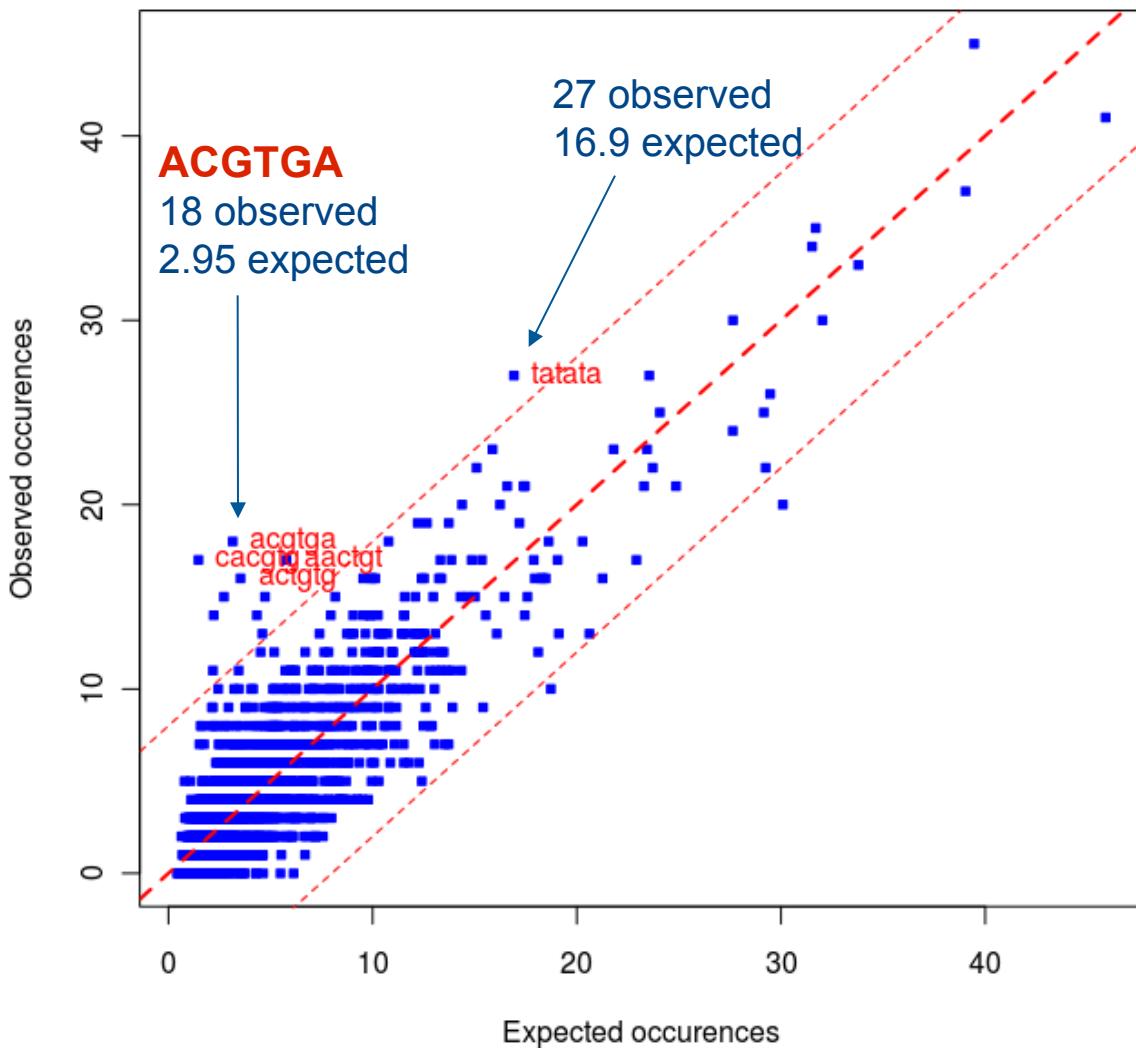
Hexanucleotide occurrences in upstream sequences of the NIT family

NIT

aaaaaaa	ttttttt	80
cttatc	gataag	26
tatata	tatata	22
ataaga	tcttat	20
aagaaaa	tttctt	20
gaaaaaa	tttttc	19
atatat	atatat	19
agataaa	ttatct	17
agaaaaa	ttttct	17
aaagaaa	ttcttt	16
aaaaca	tgtttt	16
aaaaaag	cttttt	15
agaaga	tcttct	14
tgataaa	ttatca	14
atataaa	ttatat	14



Motif discovery using word counting



How to evaluate expected number of occurrences ?

Motif discovery using word counting

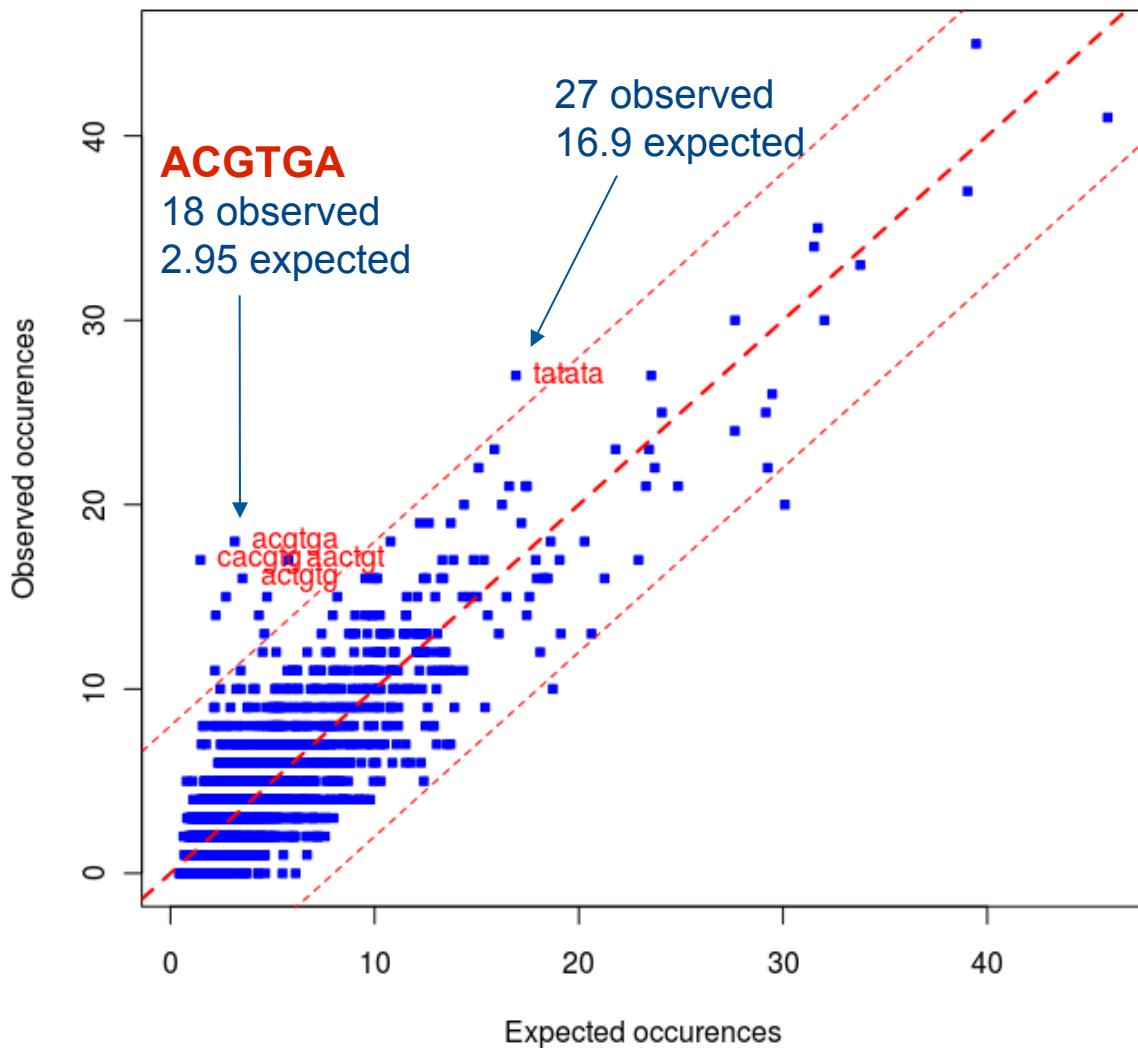
Idea:

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 evaluation



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

Statistical evaluation

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 :** $p = 3.4\text{e-}4$ (intergenic frequencies)
 $n = 9000$ position
 $x = 18$ observed occurrences

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

Binomial distribution to measure the “surprise”