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

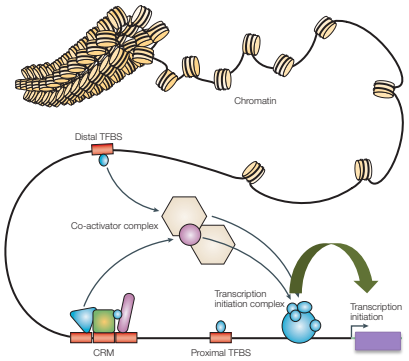
Tools

- **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*)

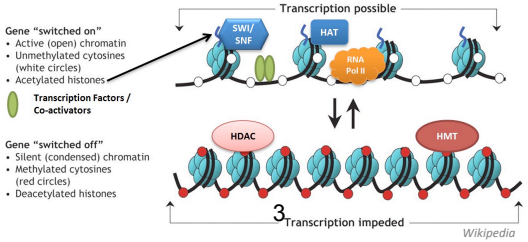
Biological concepts of transcriptional regulation



Wasserman et al, Nat Rev Genet, 2004

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

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



Wikipedia

in vivo experimental methods to identify binding sites

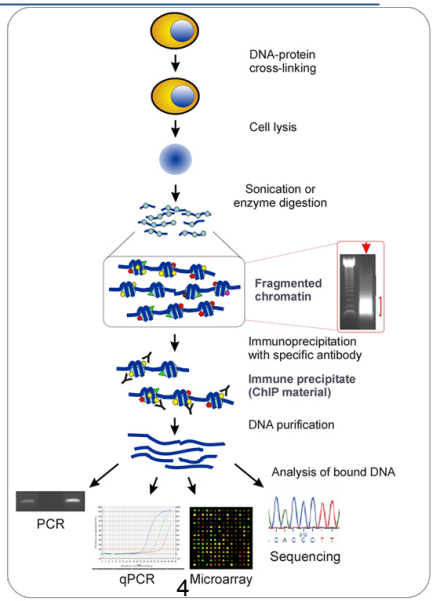
ChIP (=Chromatin Immuno-Precipitation)

=> differences in **methods to detect** the **bound DNA**

-small-scale: PCR / qPCR

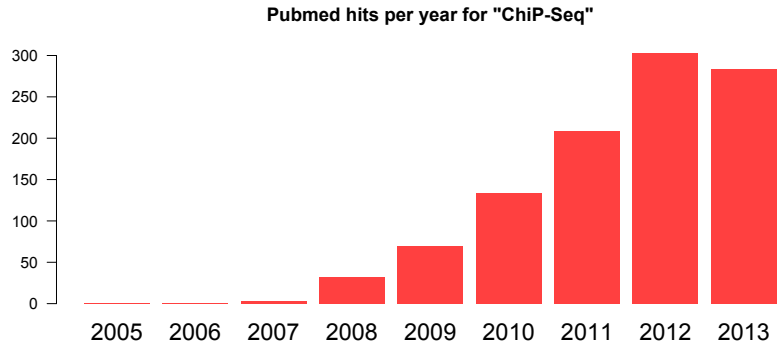
- large-scale:
  - microarray = **ChIP-on-chip**
  - sequencing = **ChIP-seq**

Main challenge:
-quality/specificity of the antibodies



http://www.chip-antibodies.com/

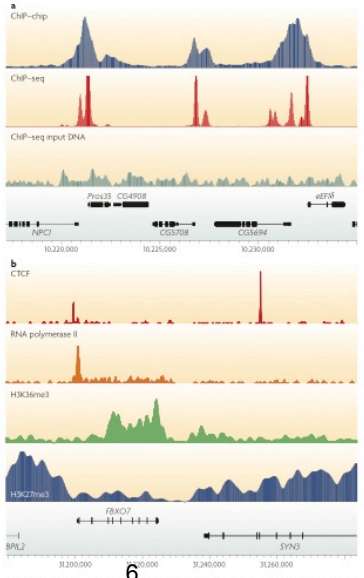
ChIP-seq is a recently-adopted technique !



ChIP-seq applications

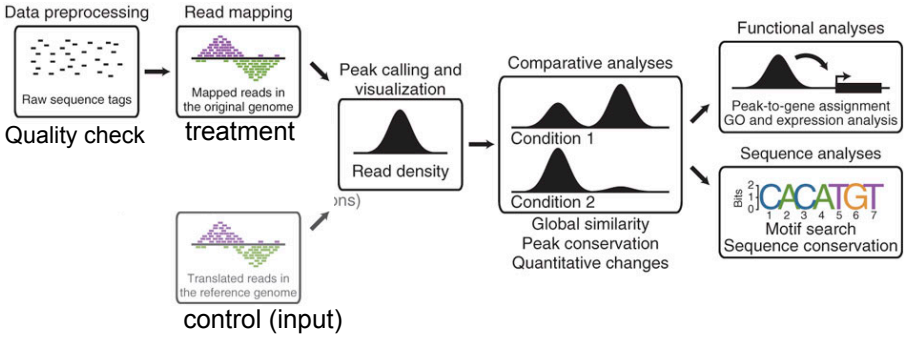
- find ***all*** regions in the genome bound by
  - a specific **transcription factor**
  - **histones** bearing a specific **modification**
- in a given ***experimental condition*** (cell type, developmental stage,...)

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



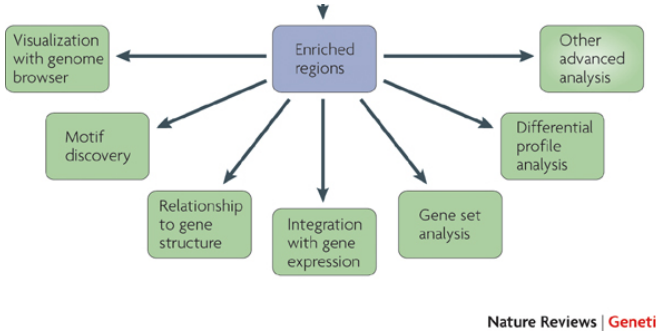
Park, Nature reviews 2009

ChIP-seq analysis workflow

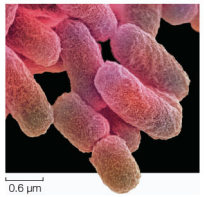


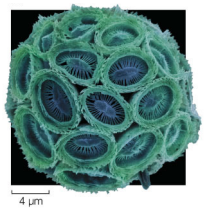
Adapted from Bardet et al, Nature Protocols, 2012

ChIP-seq analysis workflow: downstream analyses

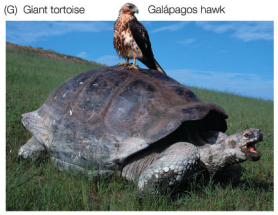


Park, Nature reviews 2009





What is the biological question ?



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

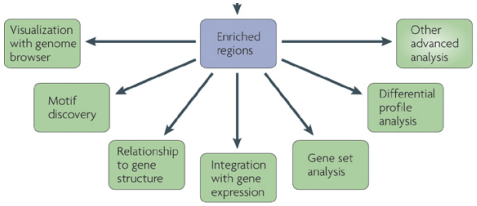
What is the biological question ?

- **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** (PolII) and **chromatin marks** ?

What is the biological question ?

➔ Should drive all « downstream » analyses

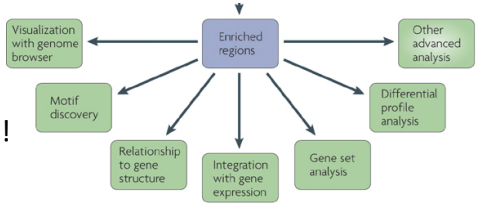


Nature Reviews | Genetics

What is the biological question ?

➔ Should drive all « downstream » analyses

Will take time to « do it all » !!!



Nature Reviews | Genetics

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

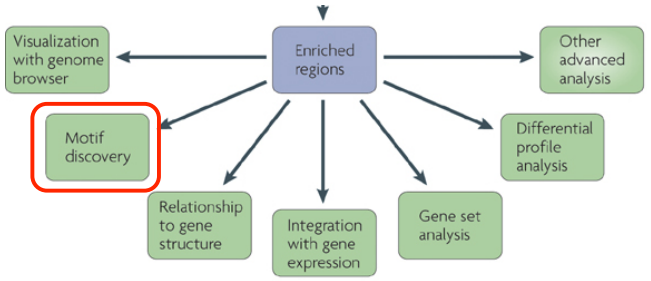
What is the biological question ?
What can be the following experimental work ?

➔ 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** ?

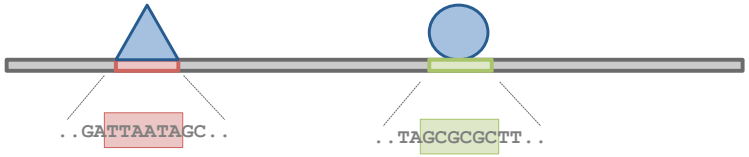
ChIP-seq analysis workflow: downstream analyses



Park, Nature reviews 2009

Transcription factor specificity

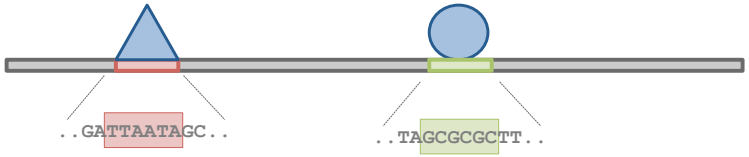
How do TF « know » where to bind DNA ?



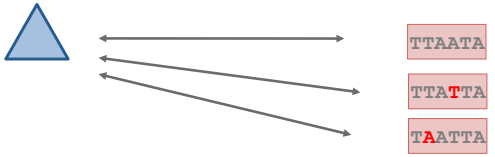
TF recognize TFBS with specific DNA sequences

Transcription factor specificity

How do TF « know » where to bind DNA ?

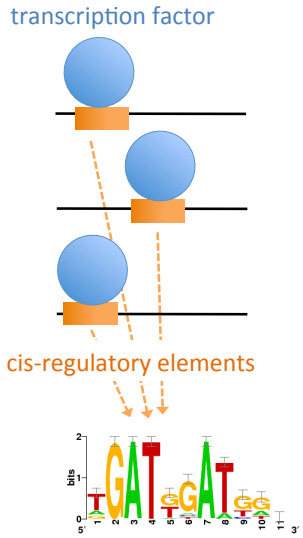


TF recognize TFBS with specific DNA sequences



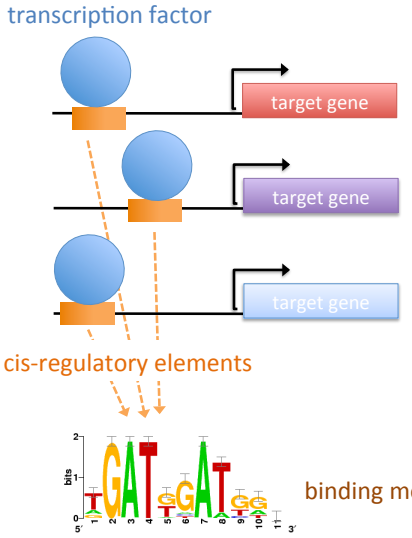
TFBSs are *degenerate*:
a given TF is able to bind DNA on TFBSs with different sequences

Binding specificity of a given TF



binding motif (represented as a sequence logo)

de novo motif discovery



binding motif

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 ?

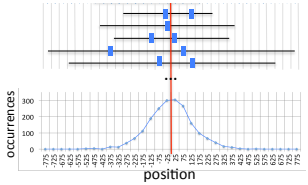
de novo motif discovery

- 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)

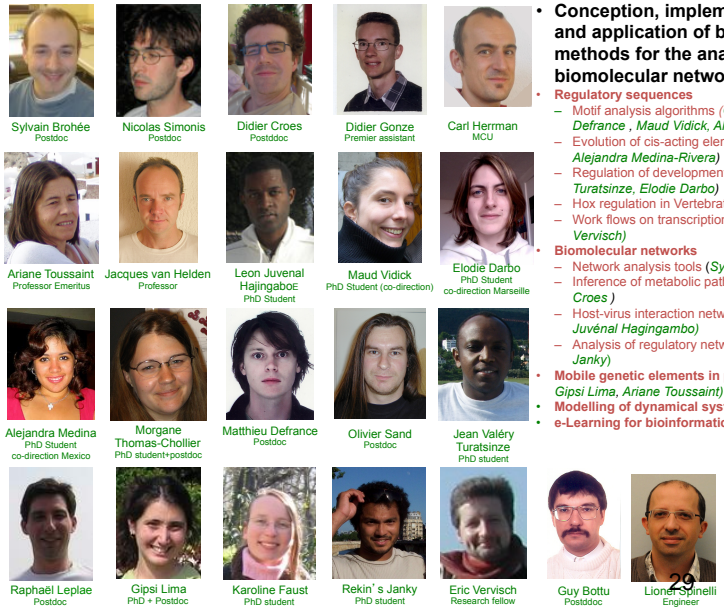
Regulatory Sequence Analysis Tools (RSAT)

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

Jacques van Helden

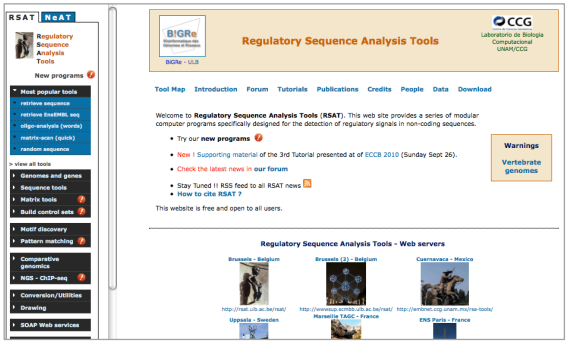
Aix*Marseille université

RSAT developers/contributors over the years



- **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)*

RSAT improvements over the years



RSAT

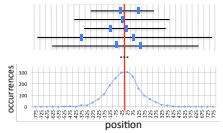
www.rsat.eu

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 !!!!!!

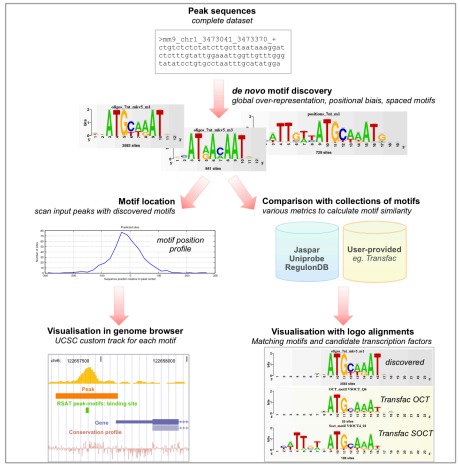
- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias



- **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

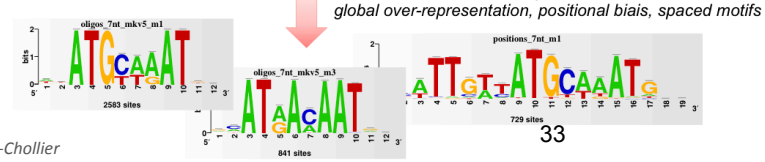
BED

```
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

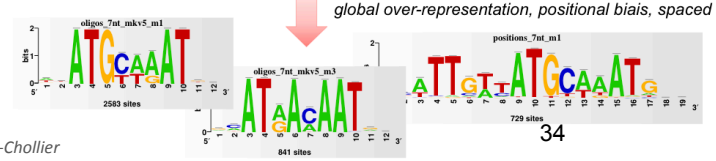
```
chr1    3002142 3002195
chr1    3002804 3002853
```

Extract corresponding sequences

Peak sequences
complete dataset

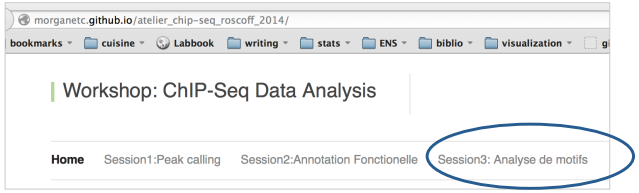
```
>mm9_chr1_3473041_3473370_+
ctgtctctctatcttgcttaataaaggat
ctctttgtattggaaattggttgtttggg
tatatcctgtgcctaatttgcatatgga
```

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



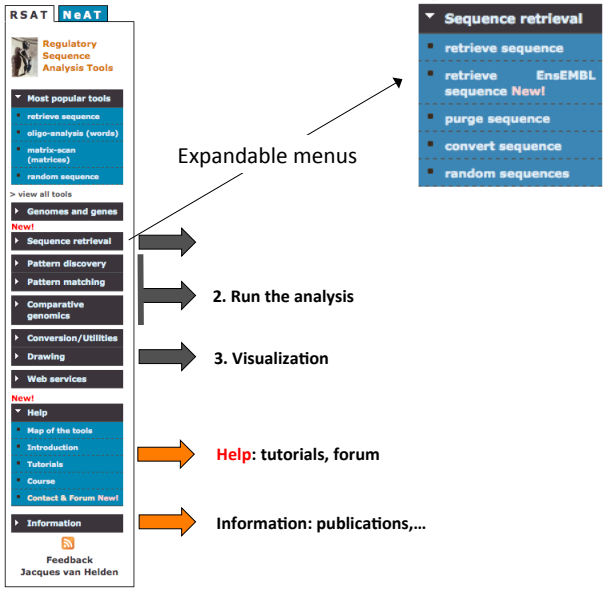
Hands on !

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

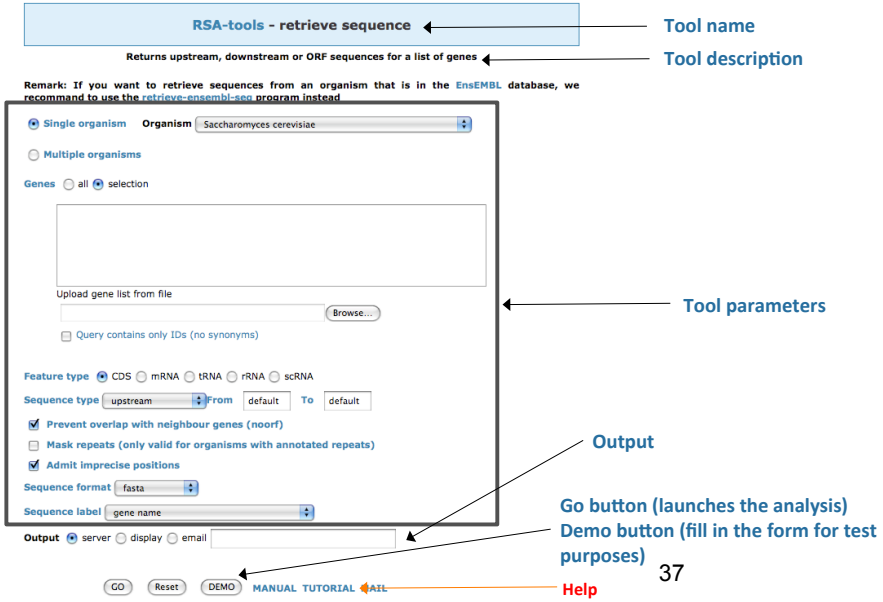


- Follow all steps of **Retrieving sequences from your peaks**

Using RSAT

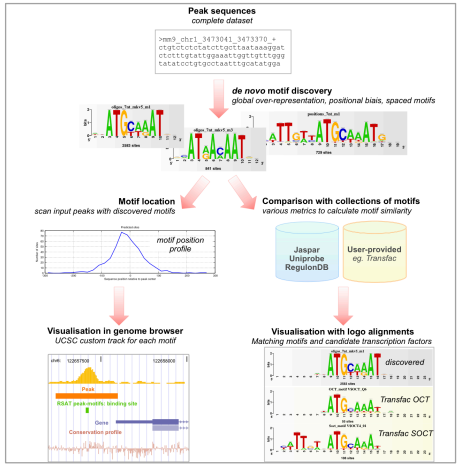


RSAT Web forms



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	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***

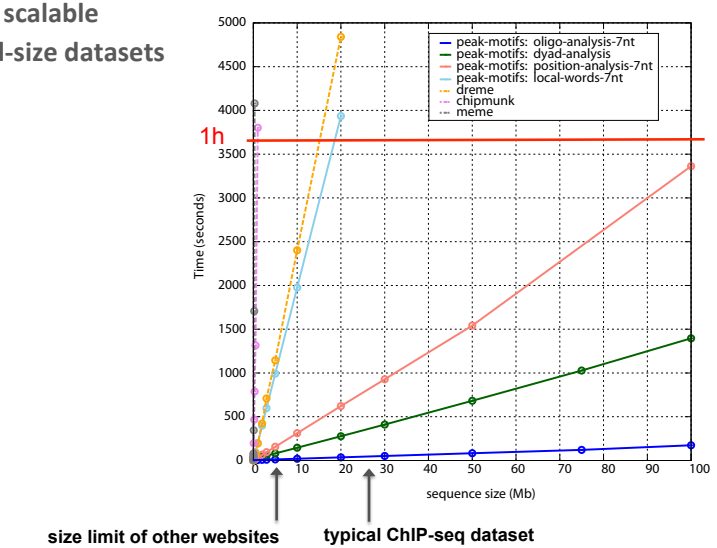
Comparison of tools for ChIP-seq

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***

Peak-motifs: why providing yet another tool ?

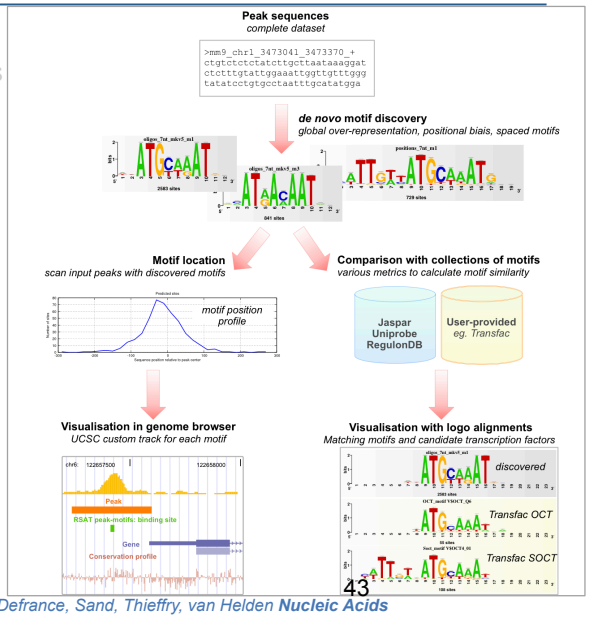
- **fast and scalable**
- **treat full-size datasets**



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

Peak-motifs: why providing yet another tool ?

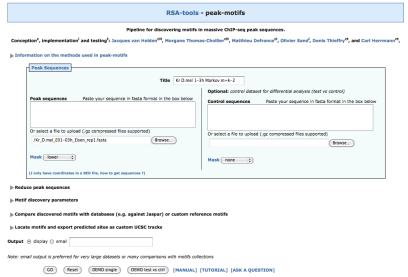
- **fast and scalable**
- **treat full-size datasets**
- **complete pipeline**



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

Peak-motifs: why providing yet another tool ?

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



- **accessible to non-specialists**
  - Demo buttons
  - Tutorials & Protocols

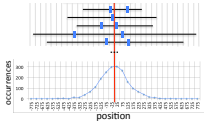
Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden ***Nature Protocols****,2012*

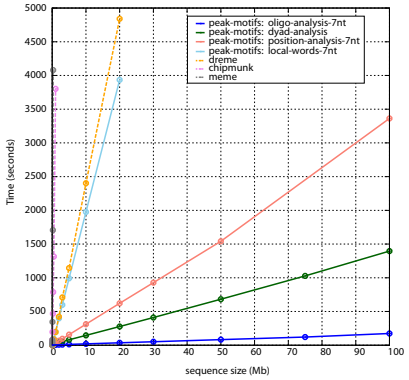
  - HTML report

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**

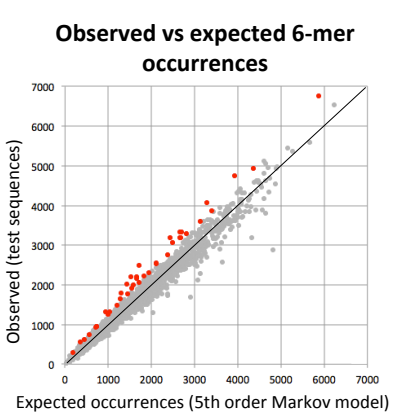




Motif discovery methods: frequency

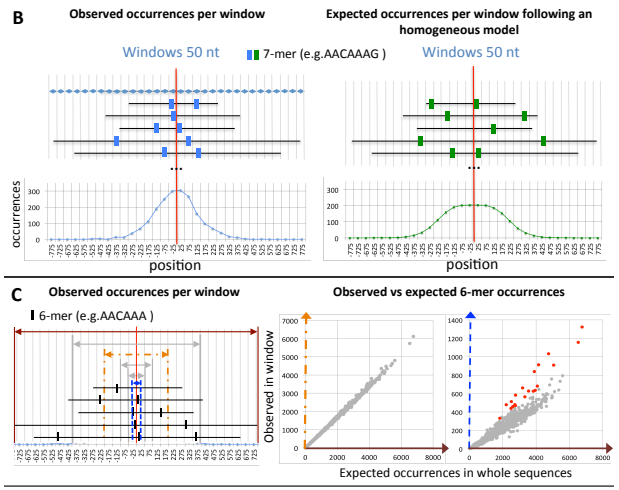
Observed 6-mer occurrences computed from:	**Expected** 6-mer occurrences computed from:
6-mer (e.g.AACAAA)	**Background sequences (when available)**
Test sequences	OR
	Theoretical k-mers frequencies from test sequences

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



oligo-analysis
dyad-analysis (spaced motifs)

Motif discovery methods: positional bias



position-analysis

local-words

Hands on !

- Go to the companion website
- Follow **all steps** 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**

Hands on !

- Go to the companion website
- Follow all steps of **Visualizing the sites in the context of genome annotations**

Hands on !

- Go to the companion website
- Follow all steps of **Motif analysis with MEME-chip**

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)

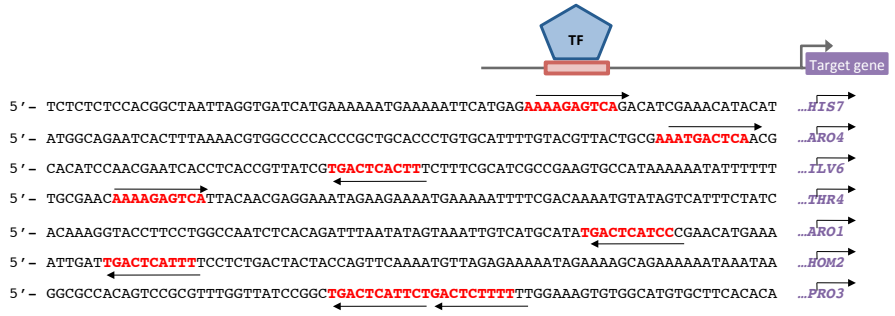
  Thomas-Chollier et al. *A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs.* Nature Protocols 7, 1551–1568 (2012).
- **Matrix-quality** : RSAT program that can be used to evaluate the enrichment of motifs in peaks

  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

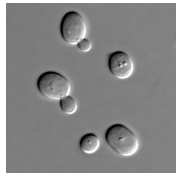
Motif discovery using word counting

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		MET		NIT	
aaaaaa|tttttt	51	aaaaaa|tttttt	105	aaaaaa|tttttt	80
aaaaag|ctttttt	15	atatat|atatat	41	**cttatc|gataag**	**26**
aagaaa|tttctt	14	gaaaaa|tttttc	40	tatata|tatata	22
gaaaaa|tttttc	13	tatata|tatata	40	ataaga|tcttat	20
tgccaa|ttggca	12	aaaaat|attttt	35	aagaaa|tttctt	20
aaaaat|attttt	12	aagaaa|tttctt	29	gaaaaa|tttttc	19
aaatta|taattt	12	agaaaa|ttttct	28	atatat|atatat	19
agaaaa|ttttct	11	aaaata|tatttt	26	agataa|ttatct	17
caagaa|ttcttg	11	aaaaag|ctttttt	25	agaaaa|ttttct	17
aaacgt|acgttt	11	agaaat|atttct	24	aaagaa|ttcttt	16
aaagaa|ttcttt	11	aaataa|ttattt	22	aaaaca|tgtttt	16
acgtgc|gcacgt	**10**	taaaaa|ttttta	21	aaaaag|ctttttt	15
aataat|attatt	10	tgaaaa|ttttca	21	agaaga|tcttct	14
aagaag|cttctt	10	ataata|tattat	20	tgataa|ttatca	14
atataa|ttatat	10	atataa|ttatat	20	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		MET		NIT	
aaaaaa|tttttt	51	aaaaaa|tttttt	105	aaaaaa|tttttt	80
aaaaag|cttttt	15	atatat|atatat	41	cttatc|gataag	26
aagaaa|tttctt	14	gaaaaa|tttttc	40	tatata|tatata	22
gaaaaa|tttttc	13	tatata|tatata	40	ataaga|tcttat	20
tgccaa|ttggca	12	aaaaat|attttt	35	aagaaa|tttctt	20
aaaaat|attttt	12	aagaaa|tttctt	29	gaaaaa|tttttc	19
aaatta|taattt	12	agaaaa|ttttct	28	atatat|atatat	19
agaaaa|ttttct	11	aaaata|tatttt	26	agataa|ttatct	17
caagaa|ttcttg	11	aaaaag|cttttt	25	agaaaa|ttttct	17
aaacgt|acgttt	11	agaaat|atttct	24	aaagaa|ttcttt	16
aaagaa|ttcttt	11	aaataa|ttattt	22	aaaaca|tgtttt	16
acgtgc|gcacgt	10	taaaaa|ttttta	21	aaaaag|cttttt	15
aataat|attatt	10	tgaaaa|ttttca	21	agaaga|tcttct	14
aagaag|cttctt	10	ataata|tattat	20	tgataa|ttatca	14
atataa|ttatat	10	atataa|ttatat	20	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)

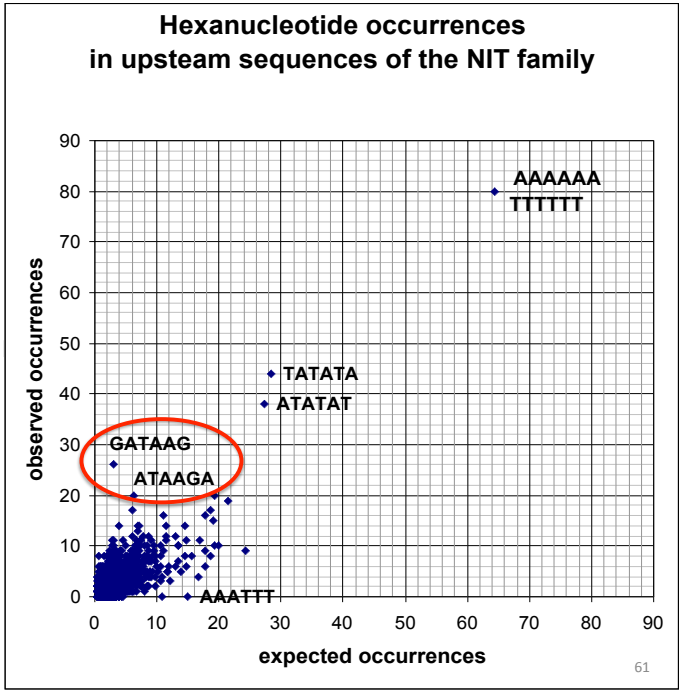
Estimation of word expected frequencies from background sequences

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

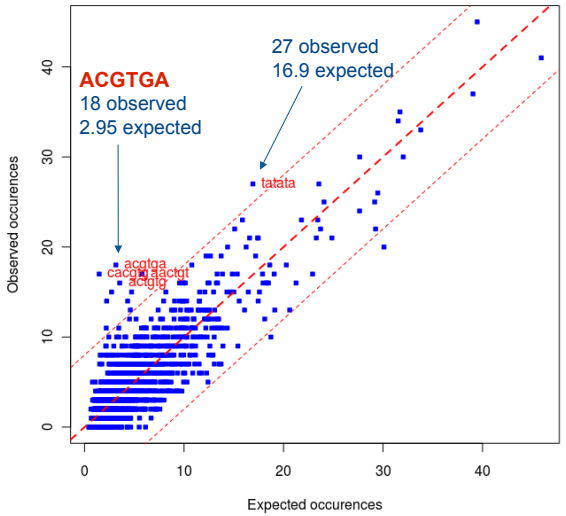
;seq	identifier	observed_freq	occ
aaaaaa	aaaaaa|ttttt	0,00510699	14555
aaaaac	aaaaac|gtttt	0,00207402	5911
aaaaag	aaaaag|ctttt	0,00375191	10693
aaaaat	aaaaat|atttt	0,00423577	12072
aaaaca	aaaaca|tgttt	0,0019828	5651
aaaacc	aaaacc|ggttt	0,00088526	2523
aaaacg	aaaacg|cgttt	0,00090105	2568
aaaact	aaaact|agttt	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

NIT

aaaaaa|tttttt	80
cttatc|gataag	**26**
tatata|tatata	22
ataaga|tcttat	**20**
aagaaa|tttctt	20
gaaaaa|ttttttc	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



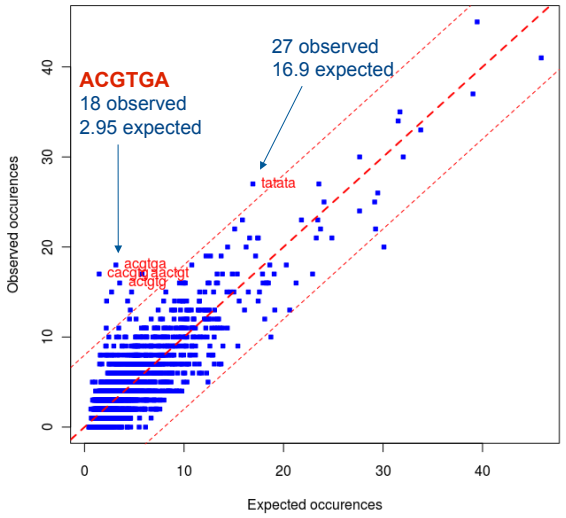
How to evaluate expected number of occurrences ?

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 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* = 3.4e-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"