TUTORIAL T01 ECCB 2014

-

Analysis of Cis-Regulatory Motifs from High-Throughput Sequence Sets

-

Session 1 : Web site

Goal and organisation of this session

Goal: introduction to ChIP-seq data analysis with RSAT

- **processing steps:** from reads to peaks. => not covered
- **downstream analyses**:
  - focus on motif analyses

Access

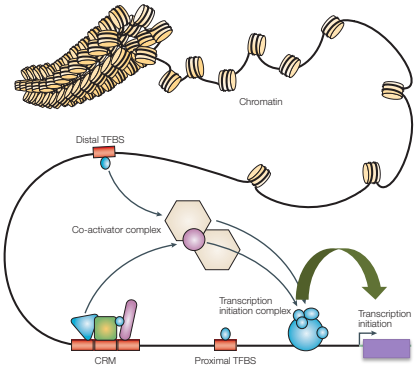
- **Web site**

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*)
4. Measure the **enrichment** of peaks for the expected motif (*matrix-quality*)
5. **A quick tour** of utility tools on RSAT

Morgane Thomas-Chollier

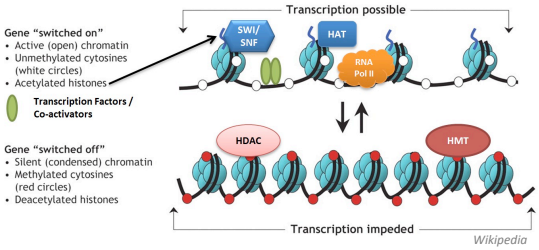
Biological concepts of transcriptional regulation



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

Morgane Thomas-Chollier

in vivo experimental methods to identify binding sites

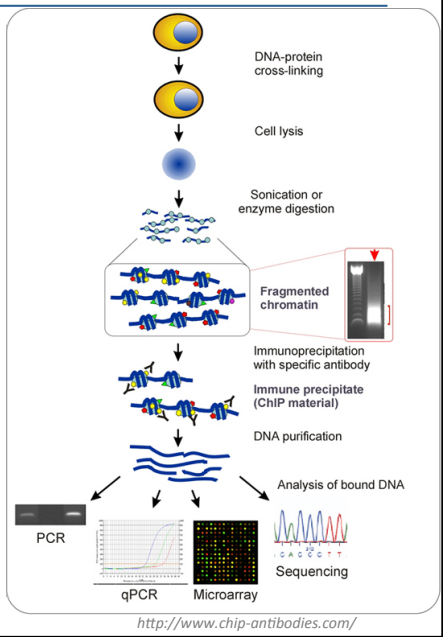
ChIP (=Chromatin Immuno-Precipitation)

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

-small-scale: PCR / qPCR

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

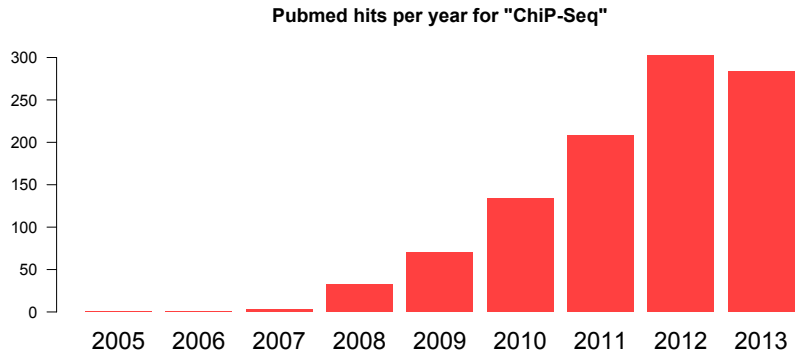
Main challenge:
-quality/specificity of the antibodies



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

Morgane Thomas-Chollier

ChIP-seq is a recently-adopted technique !

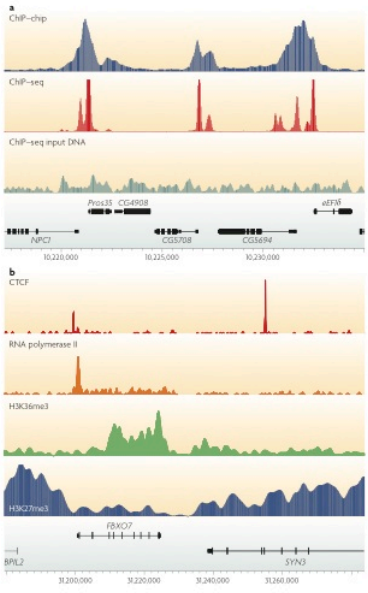


Morgane Thomas-Chollier

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,...)

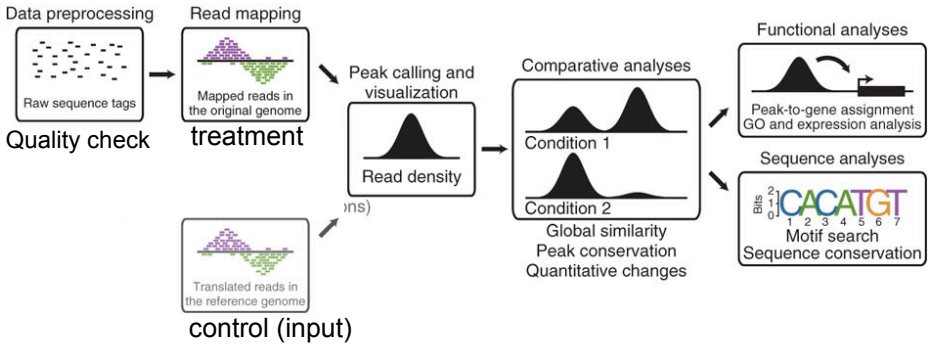
The obtain ChIP-seq **profiles** have **different shapes**, depending on the targeted protein



Park, Nature reviews 2009

Morgane Thomas-Chollier

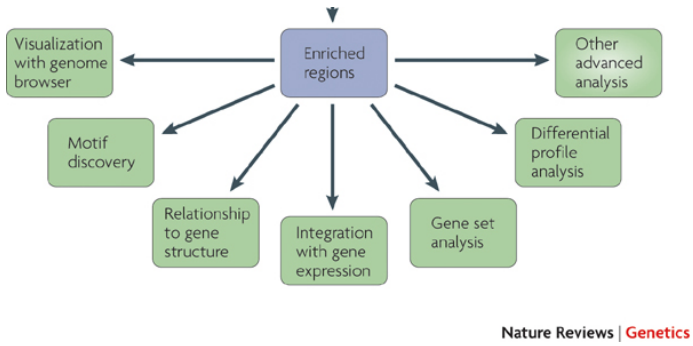
ChIP-seq analysis workflow



Morgane Thomas-Chollier

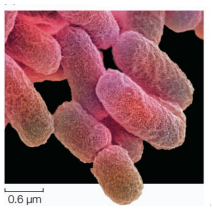
Adapted from Bardet et al, Nature Protocols, 2012

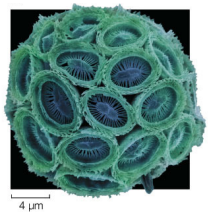
ChIP-seq analysis workflow: downstream analyses



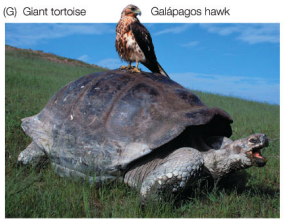
Morgane Thomas-Chollier

Park, Nature reviews 2009





What is the biological question ?



Morgane Thomas-Chollier

What is the biological question ?

« see if you can find something in the data »

Morgane Thomas-Chollier

What is the biological question ?

« see if you can find something in the data »

Morgane Thomas-Chollier

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)

Morgane Thomas-Chollier

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

Morgane Thomas-Chollier

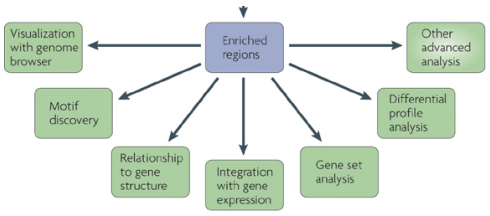
What is the biological question ?

- **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** ?

Morgane Thomas-Chollier

What is the biological question ?

➔ Should drive all « downstream » analyses



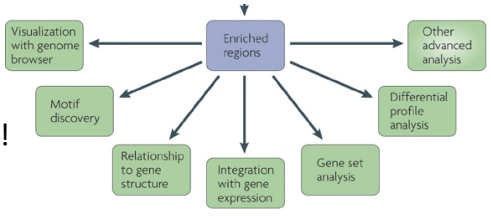
Morgane Thomas-Chollier

Nature Reviews | Genetics

What is the biological question ?

➔ Should drive all « downstream » analyses

Will take time
to « do it all » !!!



Morgane Thomas-Chollier

Nature Reviews | Genetics

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

Morgane Thomas-Chollier

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

Morgane Thomas-Chollier

What is the biological question ?

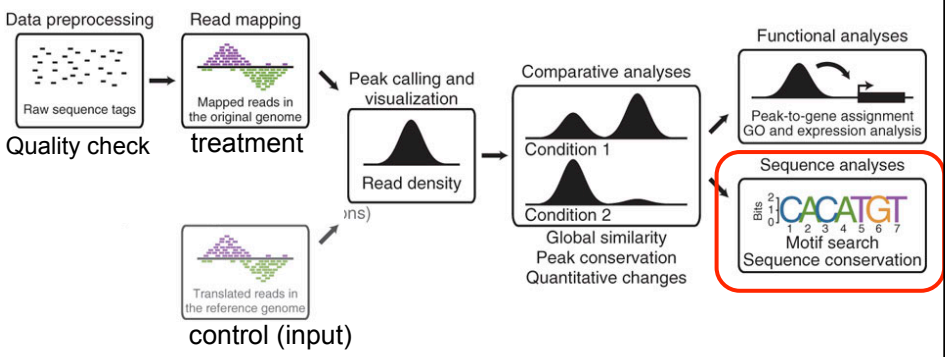
- **Where** do a transcription factor (TF) bind ?
  - ✓ In a specific context (tissue, developmental stage, mutant)
  - ✓ By comparison to another context (WT vs mutant, different time points)

- **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** ?

Morgane Thomas-Chollier

ChIP-seq analysis workflow

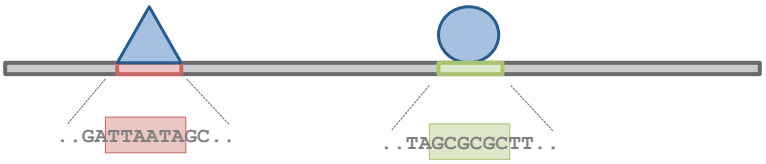


Morgane Thomas-Chollier

Adapted from Bardet et al, Nature Protocols, 2012

Transcription factor specificity

How do TF « know » where to bind DNA ?

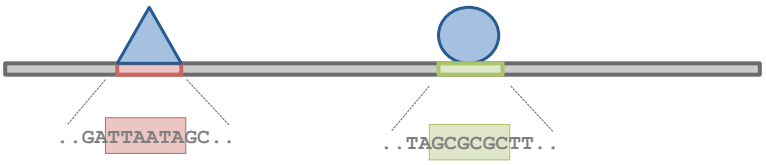


TF recognize TFBS with specific DNA sequences

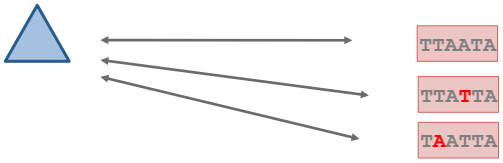
Morgane Thomas-Chollier

Transcription factor specificity

How do TF « know » where to bind DNA ?



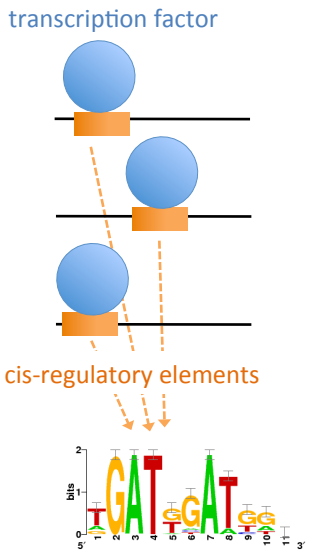
TF recognize TFBS with specific DNA sequences



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

Morgane Thomas-Chollier

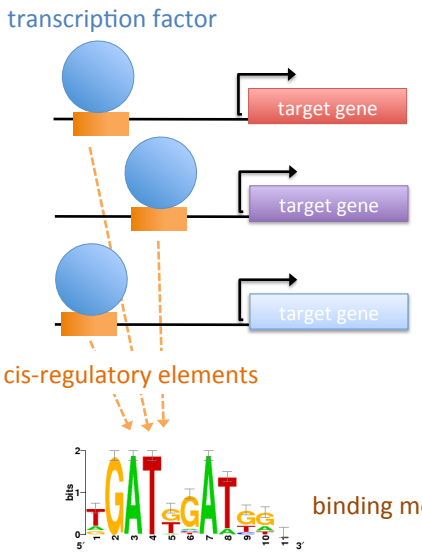
Binding specificity of a given TF



binding motif (represented as a sequence logo)

Morgane Thomas-Chollier

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 ?

Morgane Thomas-Chollier

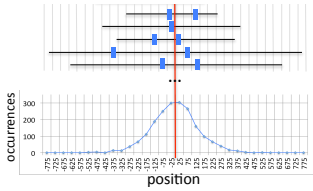
de novo motif discovery

- 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**)



Morgane Thomas-Chollier

de novo motif discovery

- Tools already exist for a long time !

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

Morgane Thomas-Chollier

de novo motif discovery

- Tools already exist for a long time !
  - MEME (1994)
  - **RSAT oligo-analysis (1998)**
  - AlignACE (2000)
  - Weeder (2001)
  - MotifSampler (2001)

Morgane Thomas-Chollier

Regulatory Sequence Analysis Tools (RSAT)

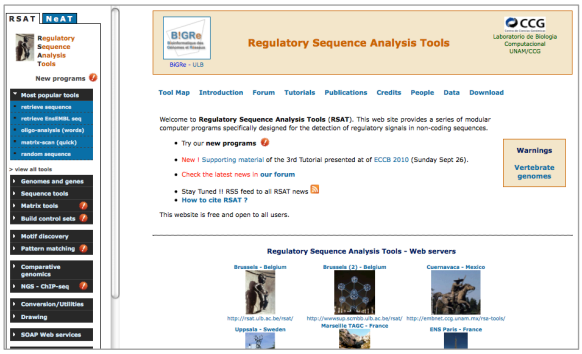
- **Since 1998 (15 years !)**
- **Initiated in Cuernavaca, Mexico**
- **yeast cis-regulatory elements**

Jacques van Helden

Aix*Marseille université

Morgane Thomas-Chollier

RSAT improvements over the years



RSAT

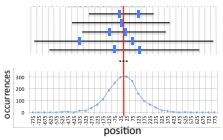
www.rsat.eu

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



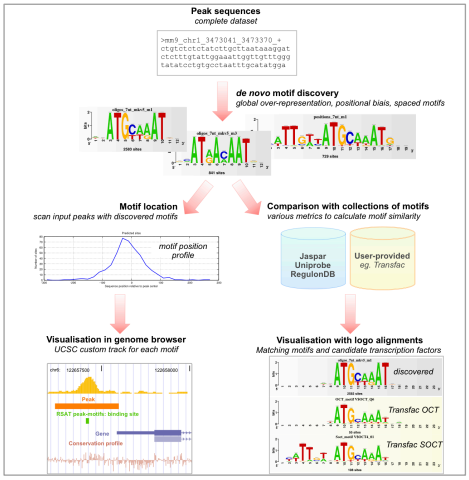
- **motif analysis: not just for specialists anymore !**
  - complete user-friendly workflows

Morgane Thomas-Chollier

http://www.genomequest.com/landing-pages/ODI-webinar-web.html

New approaches for ChIP-seq datasets

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



Thomas-Chollier et al Nucleic Acids Research, 2012

Morgane Thomas-Chollier

Peaks coordinates

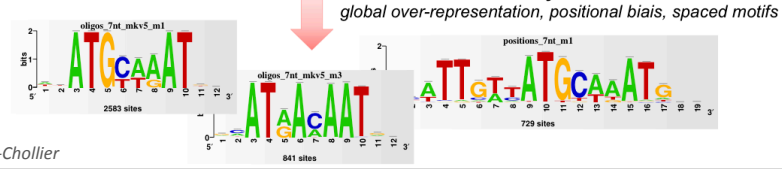
BED

```
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



Morgane Thomas-Chollier

Peaks coordinates

BED

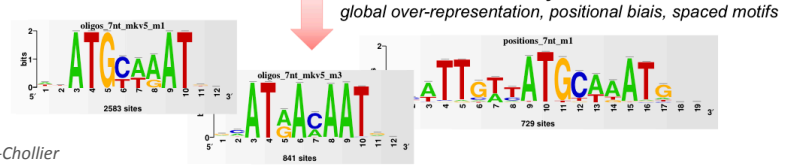
```
chr1    3002142 3002195
chr1    3002804 3002853
```

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



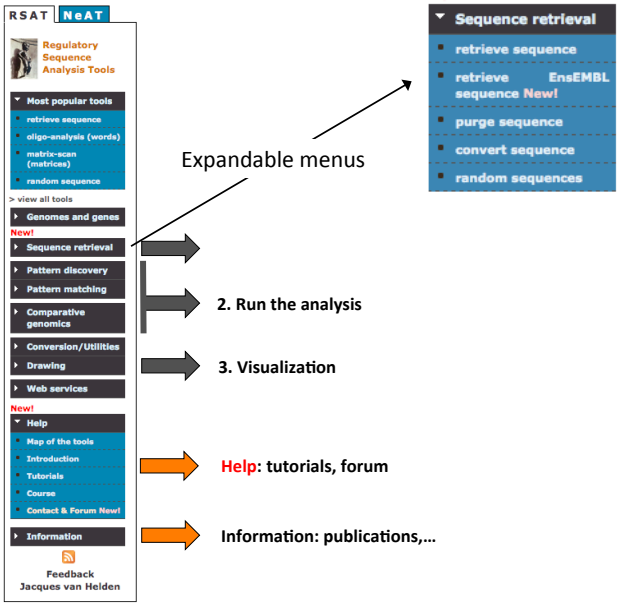
Morgane Thomas-Chollier

Hands on !

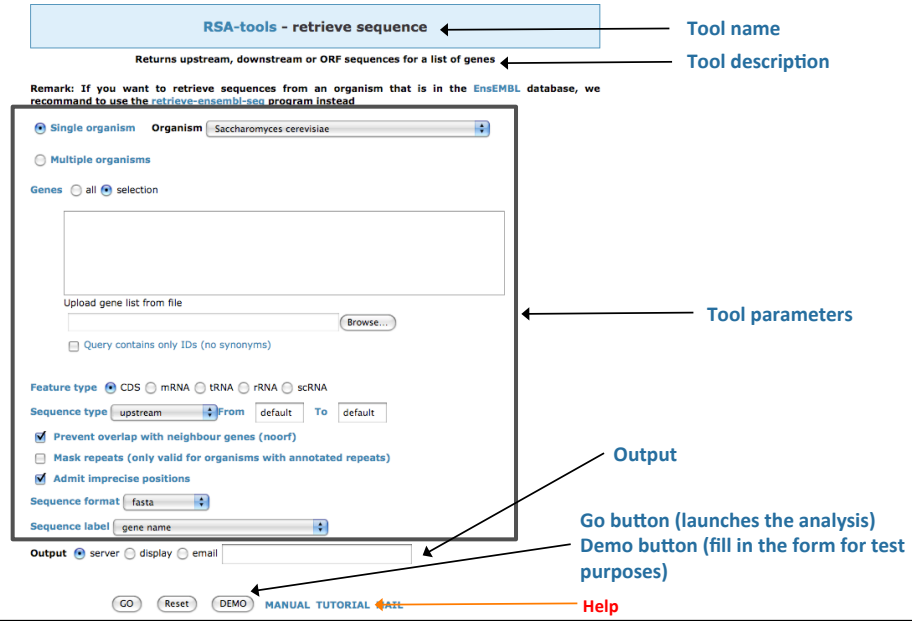
- Go to the companion website
- Follow all steps of **Retrieving sequences from your peaks**

Morgane Thomas-Chollier

Using RSAT

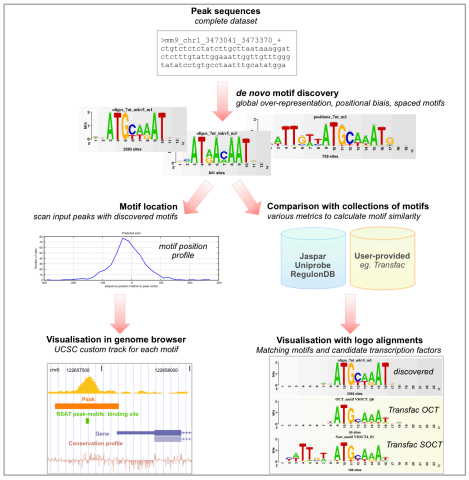


RSAT Web forms



New approaches for ChIP-seq datasets

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

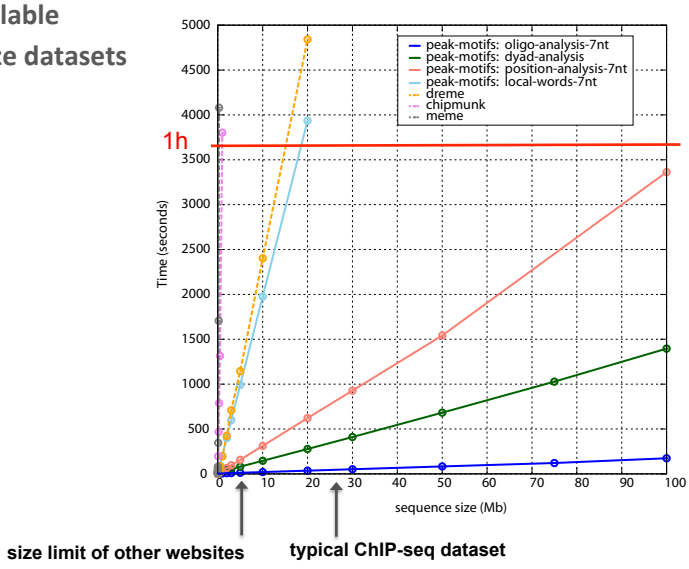


Thomas-Chollier et al Nucleic Acids Research, 2012

Morgane Thomas-Chollier

Peak-motifs: why providing yet another tool ?

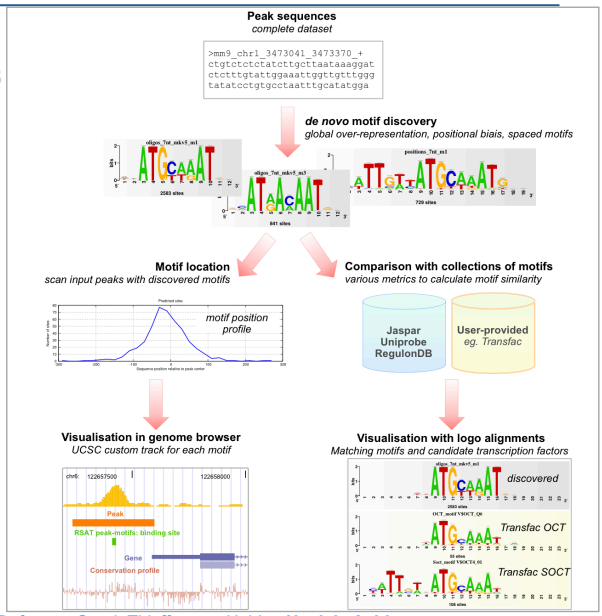
- **fast and scalable**
- **treat full-size datasets**



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

Peak-motifs: why providing yet another tool ?

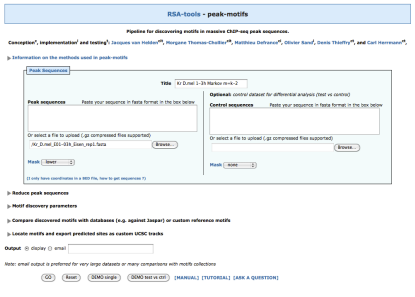
- fast and scalable
- treat full-size datasets
- complete pipeline



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

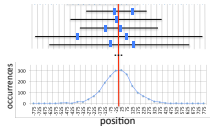
Hands on !

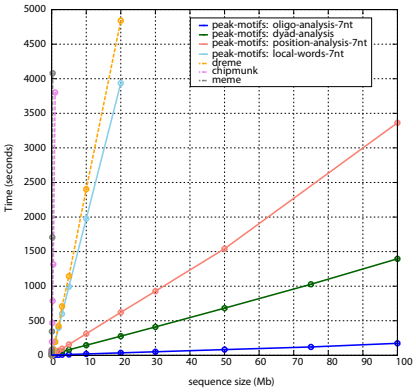
- Go to the companion website
- Follow **step 1** of **Discovering motifs from peak sequences**

Morgane Thomas-Chollier

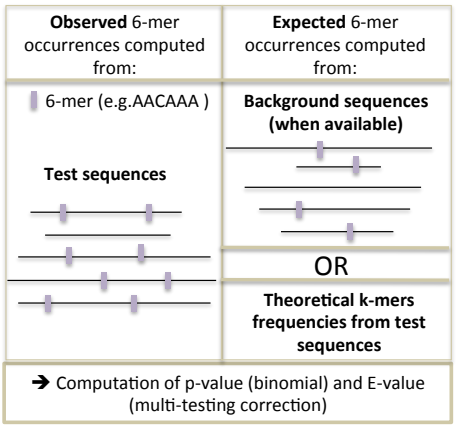
Peak-motifs: why providing yet another tool ?

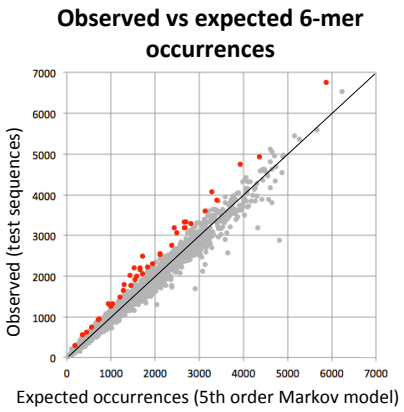
- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface
- accessible to non-specialists
- **using 4 complementary algorithms**
  - **Global over-representation**
    - **oligo-analysis**
    - **dyad-analysis (spaced motifs)**
  - **Positional bias**
    - **position-analysis**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





Motif discovery methods: frequency

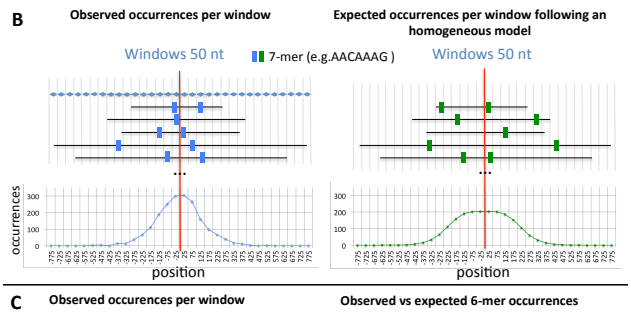




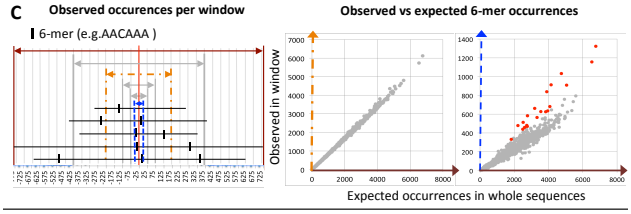
oligo-analysis
dyad-analysis (spaced motifs)

Morgane Thomas-Chollier *Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden **Nature Protocols**,20*

Motif discovery methods: positional bias



position-analysis

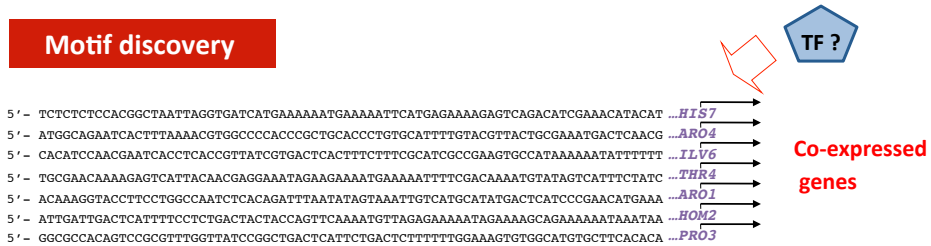


local-words

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

A motif discovery problem

Motif discovery



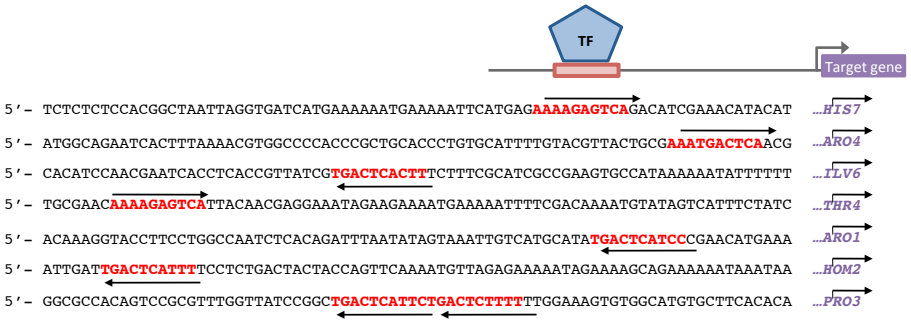
Co-expressed genes

Problem : If there is a common regulating factor, can we discover its motif (some signal) on the basis of these sequences ONLY ?

- We have a set of sequences
- We suspect that they share some functional signal
- We ignore the transcription factors involved in this regulation.
- We ignore the cis-acting elements

Morgane Thomas-Chollier

Principle: detect unexpected patterns



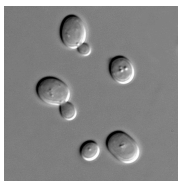
- Binding sites are represented as “words” = “string”=“k-mer”
  - e.g. **acgtga** is a 6-mer
- Signal is likely to be more frequent in the upstream regions of the co-regulated genes than in a random selection of genes
- We will thus detect over-represented words

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|cttttt	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|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|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

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|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|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|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

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)

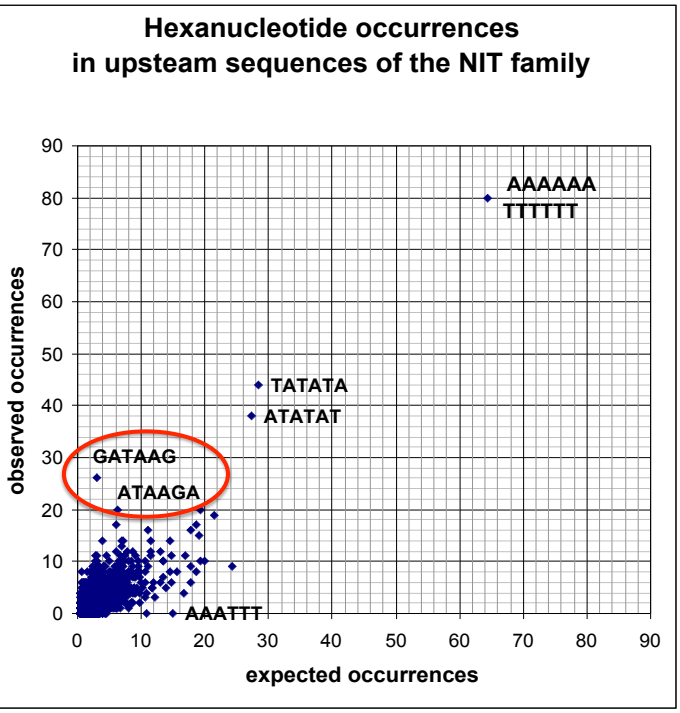
Estimation of word expected frequencies from background sequences

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

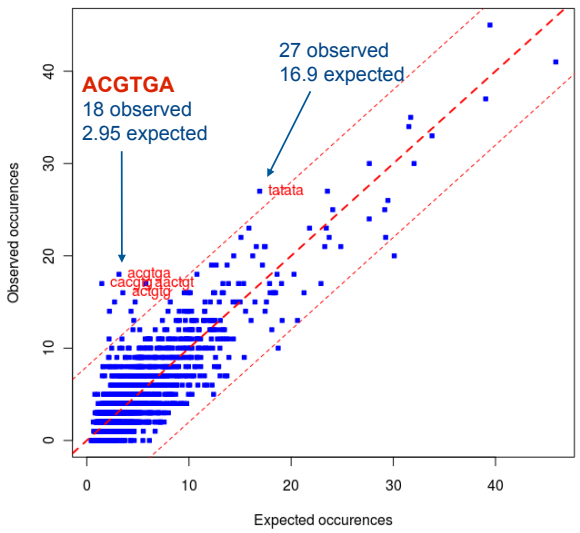
;seq	identifier	observed_freq	occ
aaaaaa	aaaaaa|ttttt	0,00510699	14555
aaaaac	aaaaac|gtttt	0,00207402	5911
aaaaag	aaaaag|ctttt	0,00375191	10693
aaaaat	aaaaat|atttt	0,00423577	12072
aaaaca	aaaaca|tgttt	0,0019828	5651
aaaacc	aaaacc|ggttt	0,00088526	2523
aaaacg	aaaacg|cgttt	0,00090105	2568
aaaact	aaaact|agttt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



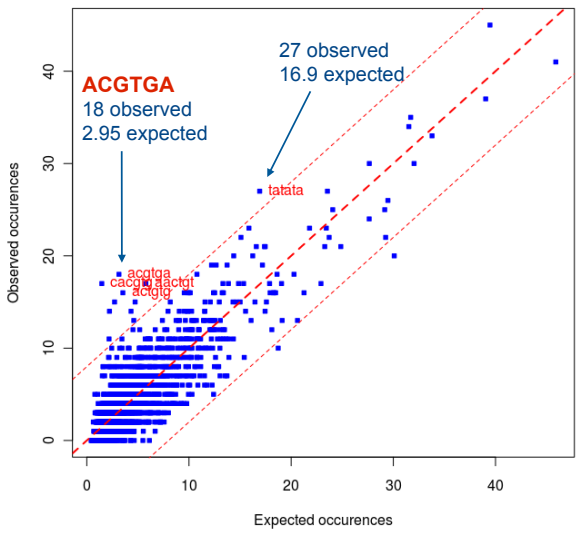
How to evaluate expected number of occurrences ?

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 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

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

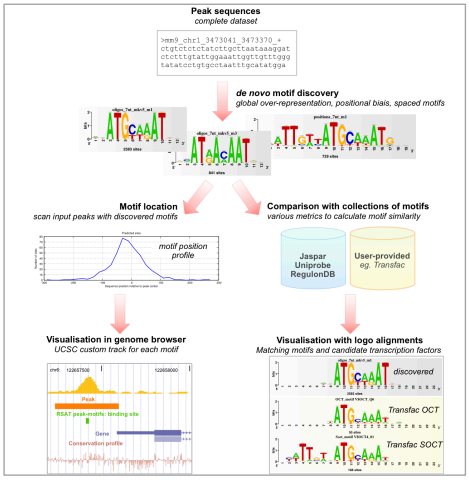
Binomial distribution to measure the "surprise"

Hands on !

- Go to the companion website
- Follow **step 2** of **Discovering motifs from peak sequences**

New approaches for ChIP-seq datasets

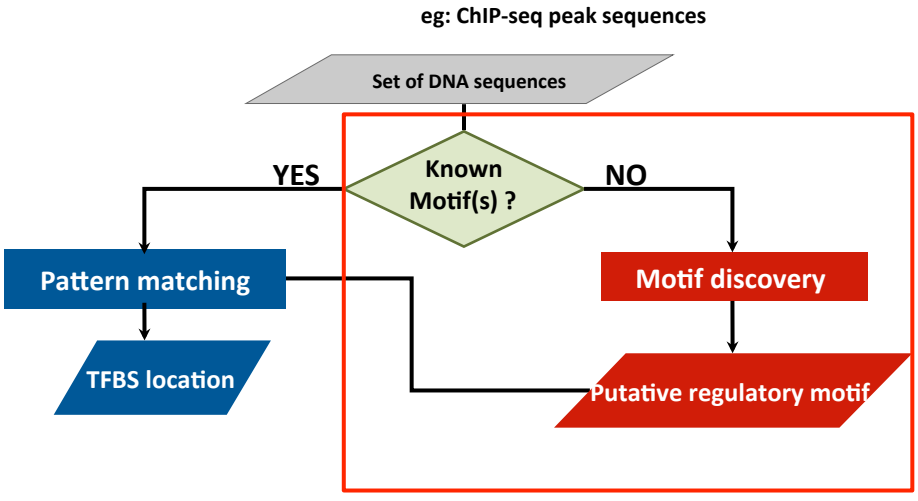
- ***de novo* motif discovery (*peak-motifs* in RSAT)**



Thomas-Chollier et al Nucleic Acids Research, 2012

Morgane Thomas-Chollier

Motif scanning vs. Motif discovery



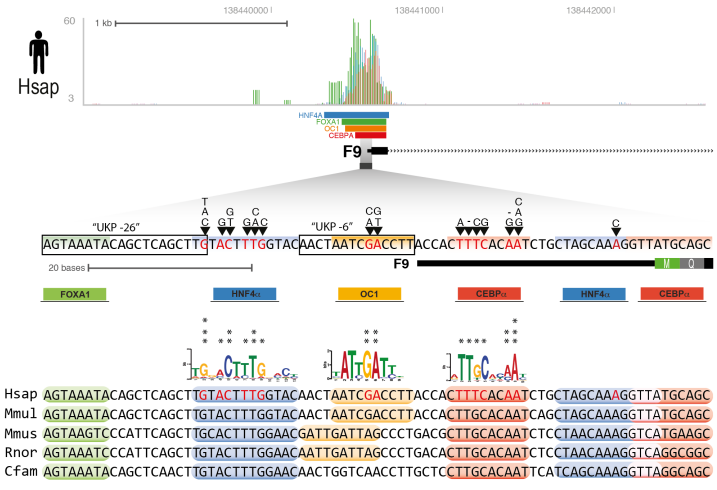
Morgane Thomas-Chollier

Hands on !

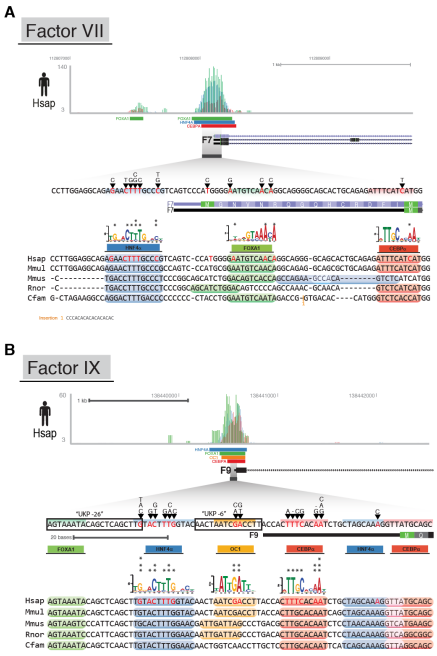
- Go to the companion website
- Follow all steps of **Visualizing the sites in the context of genome annotations**

Morgane Thomas-Chollier

Liver related genes have suffered rare mutations in their regulatory regions leading to human disease



A CpG Mutational Hotspot in a ONECUT Binding Site Accounts for the Prevalent Variant of Hemophilia B Leyden. Funnell AP, Wilson MD, Ballester B, et al. Am J Hum Genet. 2013



Ballester ,et al. in

Hands on !

- Go to the companion website
- Follow all steps of **ChIP-seq in bacteria**

Morgane Thomas-Chollier

Transcription Factor Sox2

Sox2
-Regulator in neuronal development.
-Required for both the maintenance of neural stem cells and the differentiation of specific neuron sub-types

16 Annotated binding sites in Transfac

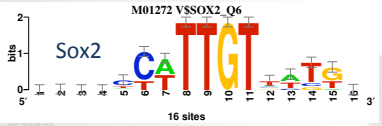
1	g	c	c	c	t	c	a	t	t	g	t	t	a	t	g	c
2	a	a	a	c	t	c	t	t	t	g	t	t	t	g	g	a
3	t	t	c	a	c	c	a	t	t	g	t	t	c	t	a	g
4	g	a	c	t	c	t	a	t	t	g	t	c	t	c	t	g
5	g	a	t	a	t	c	t	t	t	g	t	t	t	c	t	t
6	t	g	c	a	c	c	t	t	t	g	t	t	a	t	g	c
7	a	a	t	t	c	c	a	t	t	g	t	t	a	t	g	a
8	a	a	a	c	t	c	t	t	t	g	t	t	t	g	g	a
9	a	t	g	g	a	c	a	t	t	g	t	a	a	t	g	c
10	a	g	g	c	c	t	t	t	t	g	t	c	c	t	g	g
11	t	g	t	g	c	t	t	t	t	g	t	n	n	n	n	n
12	c	t	c	a	a	c	t	t	t	g	t	a	a	t	t	t
13	g	c	a	g	c	c	a	t	t	g	t	g	a	t	g	c
14	c	a	c	c	c	c	t	t	t	g	t	t	a	t	g	c
15	t	t	t	t	c	t	a	t	t	g	t	t	t	t	t	a
16	a	a	a	g	g	c	a	t	t	g	t	g	t	t	t	c

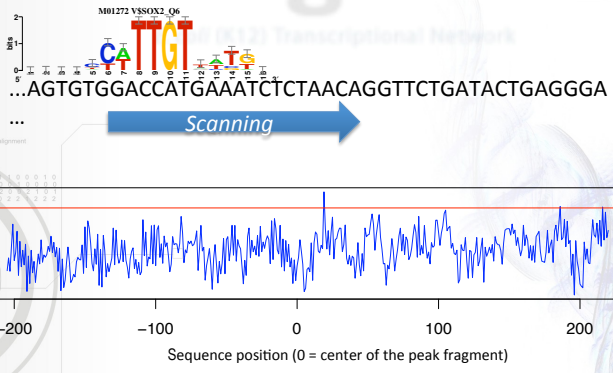
a	6	7	4	4	2	0	8	0	0	0	0	2	7	0	1	4
c	2	2	6	5	9	12	0	0	0	0	0	2	2	2	0	6
g	4	3	2	4	1	0	0	0	0	16	0	2	0	2	9	3
t	4	4	4	3	4	4	8	16	16	0	16	9	6	11	5	2

Scanning a peak sequence with a TF binding motif

UNAM

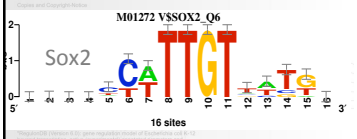
CCG Centro de Ciencias Genómicas



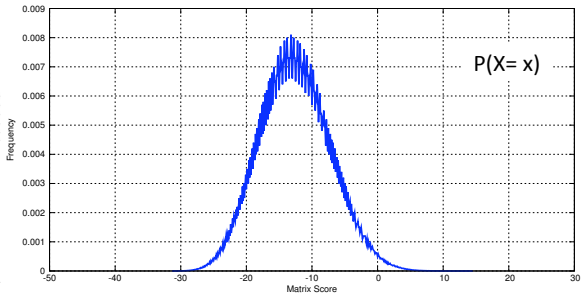


66

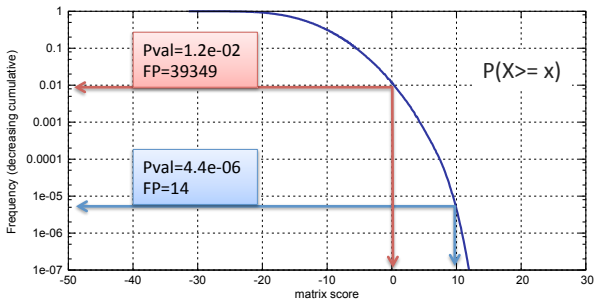
Probability density function



Theoretical Frequency of scores



Theoretical score decresing comulative distribution



Distribution of matrix scores in the previously known binding sites

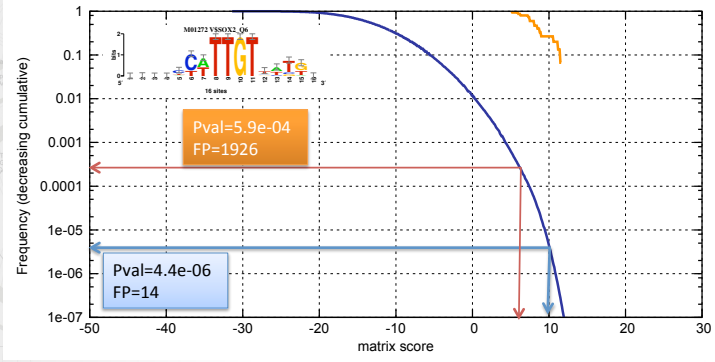
Theoretical score distribution

Matrix Sites

Distribution of matrix scores in the Sox2 binding sites annotated in TRANSFAC

TGCACCTTTGTTATGC 11.6

AGGCCTTTTGTCCTGG 5.1



Unbiased estimation of matrix scores in known binding sites

Theoretical score distribution

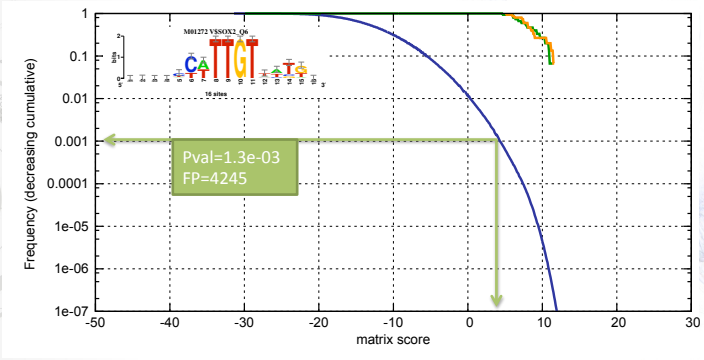
Matrix Sites

Matrix Sites LOO

- Control Over-fitting bias
- Measure how good is the matrix on detecting sites it doesn't know

TGCACCTTTGTTATGC 10.9514

AGGCCTTTTGTCCTGG 4.65353

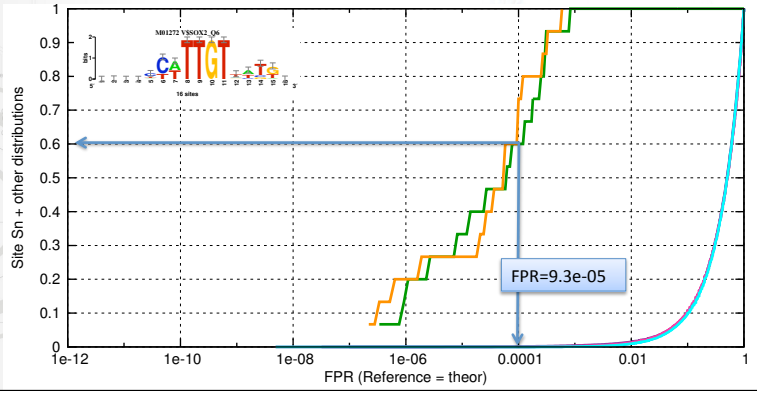


Matrix Quality False Positive Rate

Matrix Sites

Matrix Sites LOO

ROC curves show which is the expected false positive rate for a given sensitivity value.



Empirical distribution of matrix scores in peak regions

- Analysis of ChIP –seq peaks
- Evaluation how "good" is the peak collection.
- Test if the collection is enriched in putative binding sites.

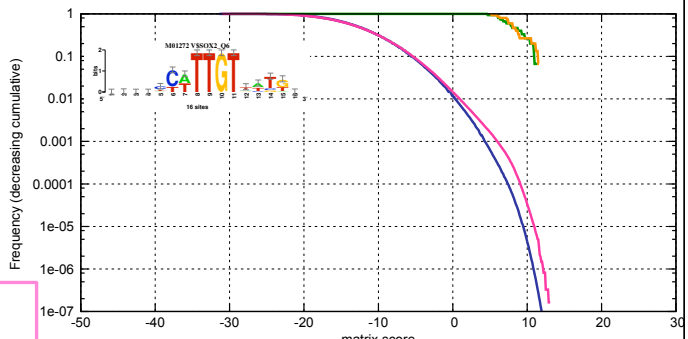
Theoretical score distribution

Matrix Sites

Matrix Sites LOO

Peaks called by MACS*

Enrichment=
High frequency of high-scoring sites than



*Data from Thomas-Chollier

71

Negative control: empirical distribution of scores for randomized matrices

Theoretical score distribution

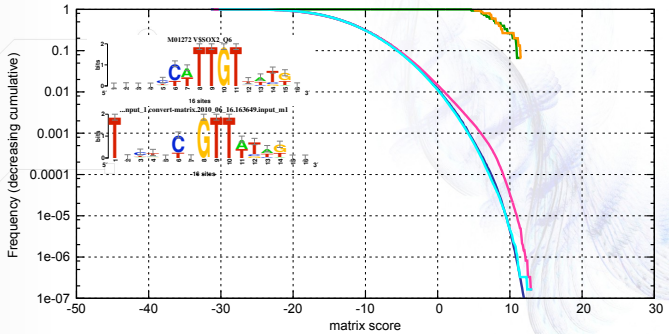
Matrix Sites

Matrix Sites LOO

Peaks called by MACS*

Column-permuted matrix in peaks called by MACS*

- Analysis of ChIP –seq data.
- Empirical negative control : scan the same sequence set (peak sequences) with randomized motifs (column-permuted matrices).



72

*Datos de Thomas-Chollier

Comparando picos de diferentes fuentes

Theoretical score distribution

Matrix Sites

Matrix Sites LOO

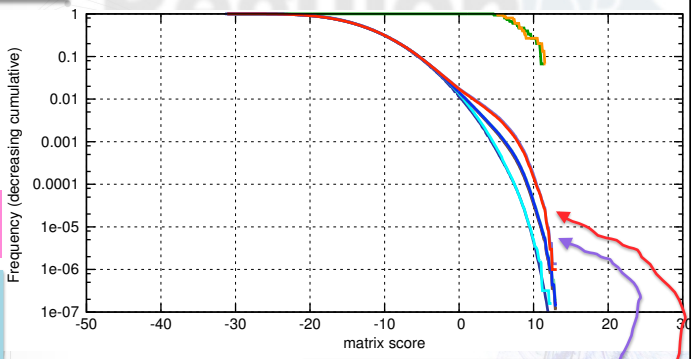
Peaks from Jaspar

Peaks called by MACS fdr 0.02*

Permuted matrix in Peaks called by MACS fdr 0.02*

Peaks called by MACS no filter.*

*Data from Thomas-



Split Peaks fdr 0.02*

Split Peaks fdr 0.02 60bps around center*

Split Peaks no filter*

Split Peaks no filter 60bps around center*

73

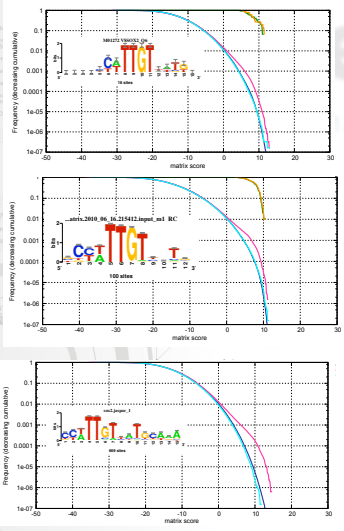
Comparación de diferentes matrices

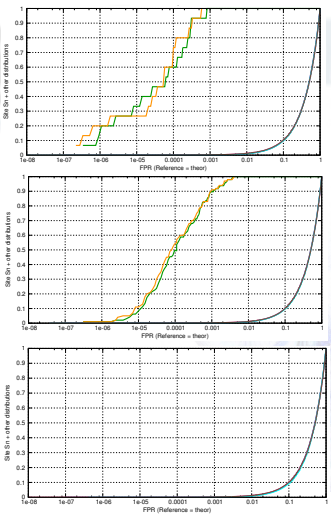
Alternative Matrices for Sox2

Transfac

Position-specific priors MEME Bailey (2010)

MEME from Chen (2008) by Jaspar team





74

Hands on !

- Go to the companion website
- Follow all steps of **Measure the enrichment of your peak for the expected motif**