Hands-on introduction to ChIP-Seq analysis

Morgane Thomas-Chollier
mthomas@biologie.ens.fr

Computational Systems Biology
*Institut de Biologie de l'Ecole Normale Supérieure, **Paris, France***

VIB Bioinformatics Training – Leuven (Belgium) – 1st June 2015

Goal and organisation of the day

Goal: introduction to ChIP-seq data analysis

- **processing steps:** from reads to peaks.
- **downstream analyses**:
  - deciding which downstream analyses to perform depending on the biological question.
  - focus on motif analyses

Schedule

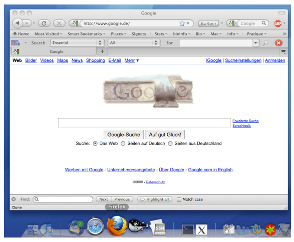
09h30-10h00 Short introduction, computer warm-up, overview of the analyses
10h00-12h30 Hands-on training: processing steps

LUNCH ☺

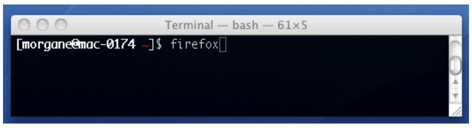
13h15-15h15 Hands-on training: downstream analysis: motifs
15h30-17h00 Discussion, feedback and questions

Don't hesitate to ask questions ☺

Why will we use the command-line ?

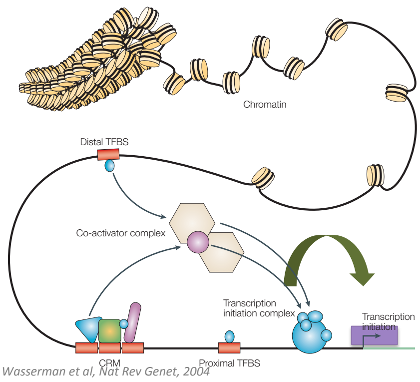


- To use a program, you usually click on the program's icon. e.g. Firefox
- The command-line is the « secret backdoor » to use a program. You need **a shell (= Terminal)** and type the name of the program you want to launch in it:



- **Why is it useful** (and mandatory sometimes !):
  - Some programs can only be run from the command-line (no icon for them)
  - When you want to use a program that is not directly installed on your machine. You can connect to a remote machine via the terminal, and run the program there.
  - To run the same program 1000 times, you might not want to click on the icon 1000 times. Instead, you can write a short program that will automatically run its command-line 1000 times.

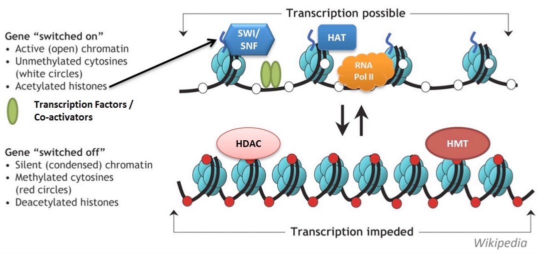
Biological concepts of transcriptional regulation



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

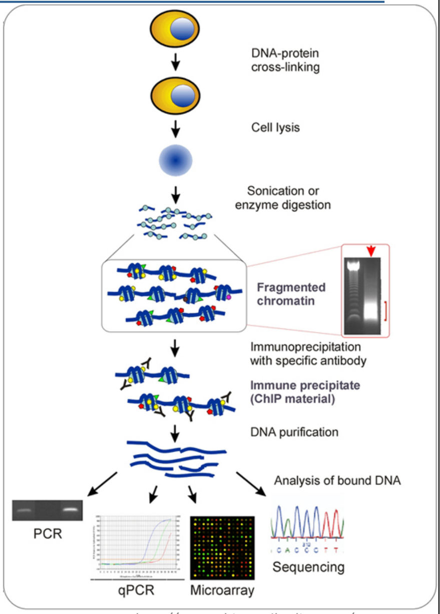
ChIP (=Chromatin Immuno-Precipitation)

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

-small-scale: PCR / qPCR

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

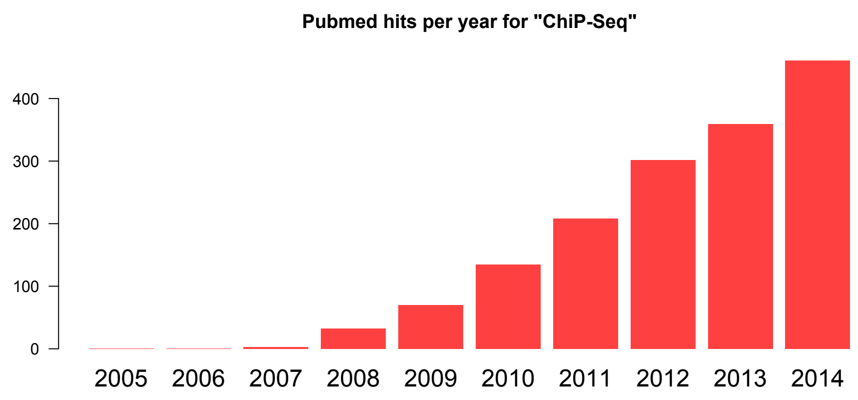
Main challenge:
-quality/specificity of the antibodies



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

Morgane Thomas-Chollier

ChIP-seq is a recently-adopted technique !

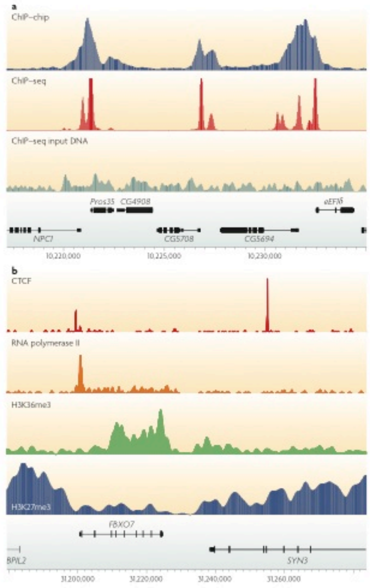


Morgane Thomas-Chollier

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,...)

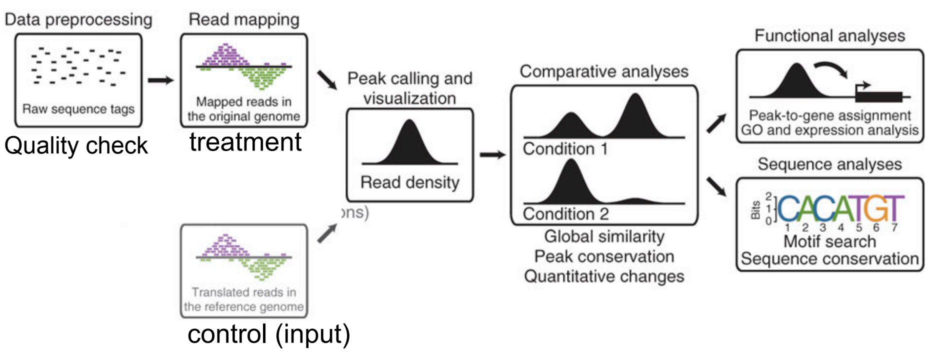
The obtain ChIP-seq **profiles** have **different shapes**, depending on the targeted protein



Morgane Thomas-Chollier

Park, Nature reviews 2009

ChIP-seq analysis workflow



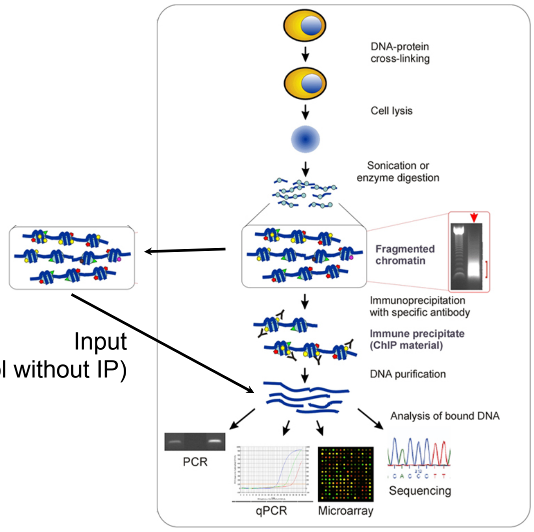
Processing steps

Downstream analyses

Morgane Thomas-Chollier

Adapted from Bardet et al, Nature Protocols, 2012

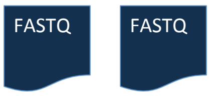
Control: input



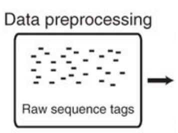
Input
(control without IP)

From sequence reads to peaks

experiment Input



sequences (reads length 36 / 50 bp, single-end)
from Illumina



Quality check

FASTQ format 1 read = 4 lines FASTA format

```
@SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
+
IIIIG1?II;IIIIII1IIII1%.I7I
@SRR002012.2 Oct4:5:1:804:348
GTCTGCATTATCTACCAGCACTTCCC
+
IIIIIIIII'I2IIIII:)I2II3I0
@SRR002012.3 Oct4:5:1:767:334
GCTGTCTTCCCGCTGTTTTATCCCCC
+
III8IIIIIII3III6II%II*III3
@SRR002012.4 Oct4:5:1:805:329
GTAGTTTACCTGTTCATATGTTTCTG
+
IIIIIII9IIIIII?IIIIIIII7II
```

```
>SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
>SRR002012.2 Oct4:5:1:804:348
GTCTGCATTATCTACCAGCACTTCCC
>SRR002012.3 Oct4:5:1:767:334
GCTGTCTTCCCGCTGTTTTATCCCCC
>SRR002012.4 Oct4:5:1:805:329
GTAGTTTACCTGTTCATATGTTTCTG
```

adapted from Wikipedia

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0                                       40

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
```

Hands on !

- Go to the companion website
- Read the **introduction**
- Follow all steps of **Downloading ChIP-seq reads from NCBI**

Morgane Thomas-Chollier

From sequence reads to peaks

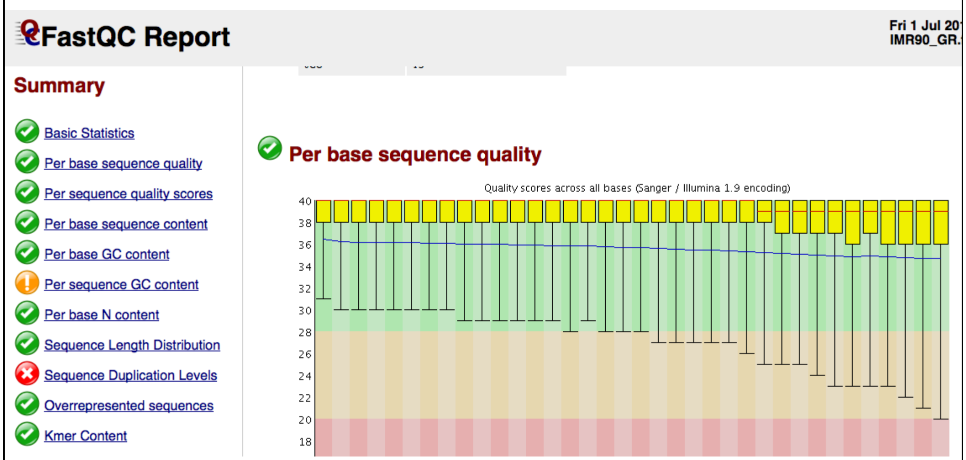
experiment Input

FASTQ FASTQ

sequences (reads length 36 / 50 bp, single-end)
from Illumina

FASTQC → **quality check**

Morgane Thomas-Chollier



http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Hands on !

- Go to the companion website

http://www.biologie.ens.fr/~mthomas/other/chip-seq-training/index.html

- Follow all steps of **Quality control of the reads and statistics**

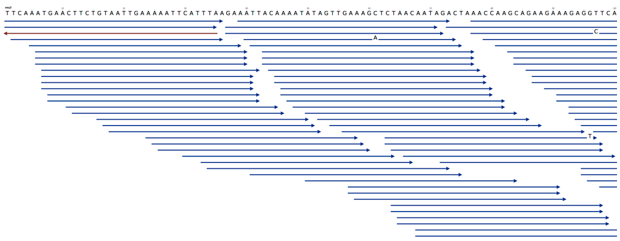
Morgane Thomas-Chollier

From sequence reads to peaks

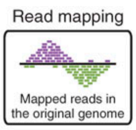
experiment Input

FASTQ FASTQ

FASTQC



Source: http://trac.seqan.de



FASTQ FASTQ

mapping

Bowtie
BWA

BED
BAM
SAM

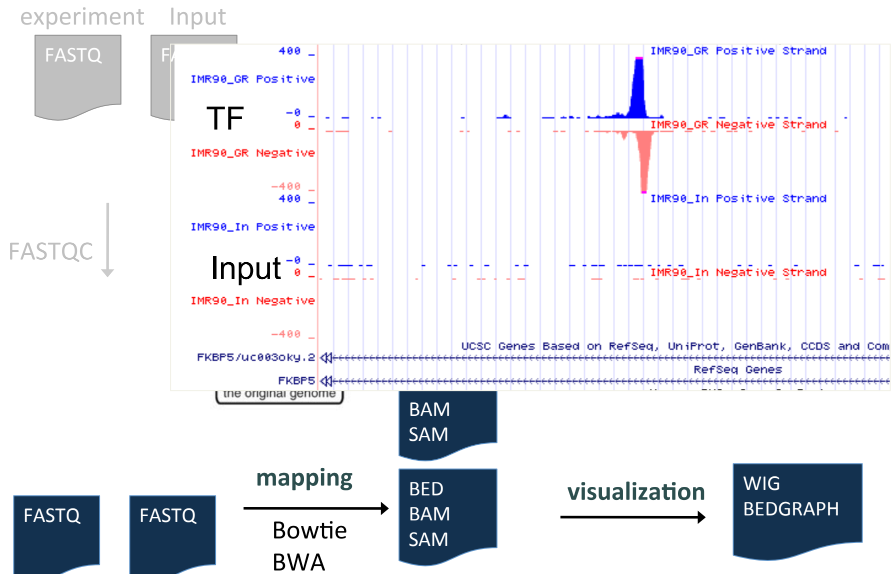
BED
BAM
SAM

Hands on !

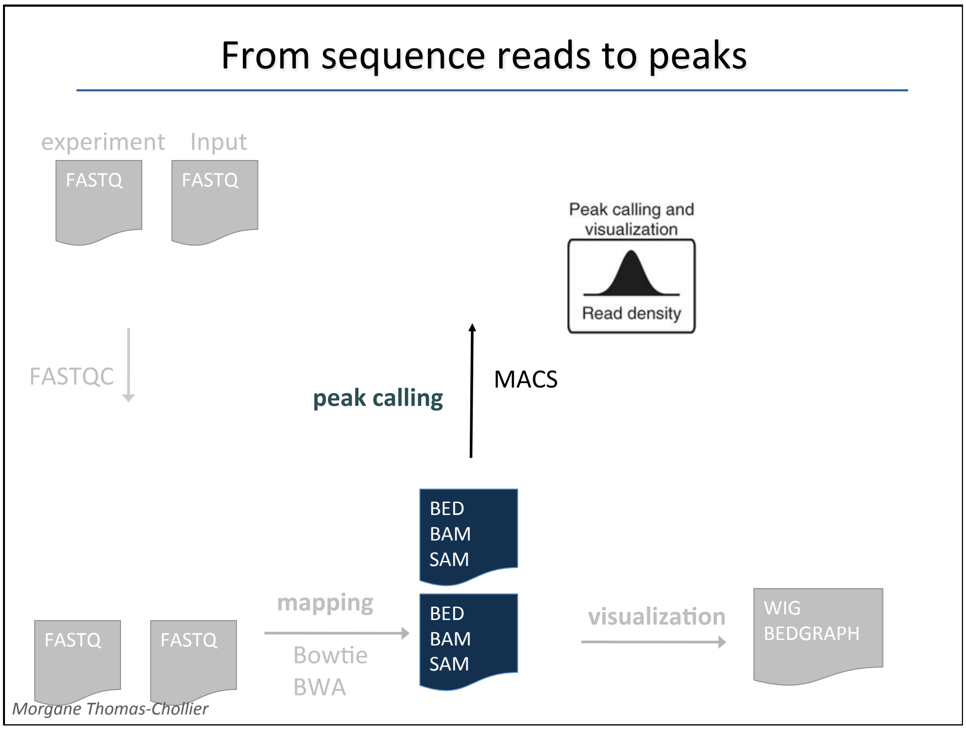
- Go to the companion website
- Follow all steps of **Mapping the reads with Bowtie**

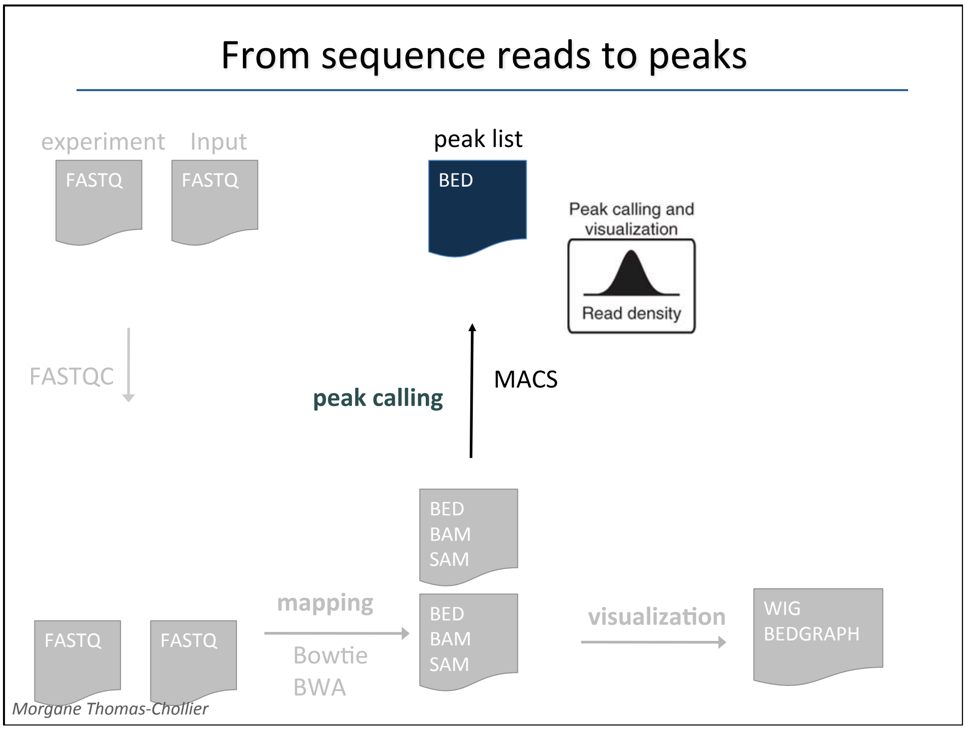
Morgane Thomas-Chollier

From sequence reads to peaks



Morgane Thomas-Chollier


From sequence reads to peaks
experiment
Input
FASTQ
FASTQ
Peak calling and visualization
Read density
FASTQC
peak calling
MACS
BED
BAM
SAM
mapping
BED
BAM
SAM
visualization
WIG
BEDGRAPH
FASTQ
FASTQ
Bowtie
BWA
Morgane Thomas-Chollier

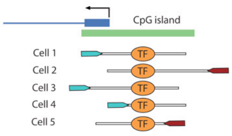

From sequence reads to peaks
experiment
Input
peak list
FASTQ
FASTQ
BED
Peak calling and visualization
Read density
FASTQC
peak calling
MACS
BED
BAM
SAM
mapping
BED
BAM
SAM
visualization
WIG
BEDGRAPH
FASTQ
FASTQ
Bowtie
BWA
Morgane Thomas-Chollier

Hands on !

- Go to the companion website
- Follow all steps of **Peak calling with MACS**

Morgane Thomas-Chollier

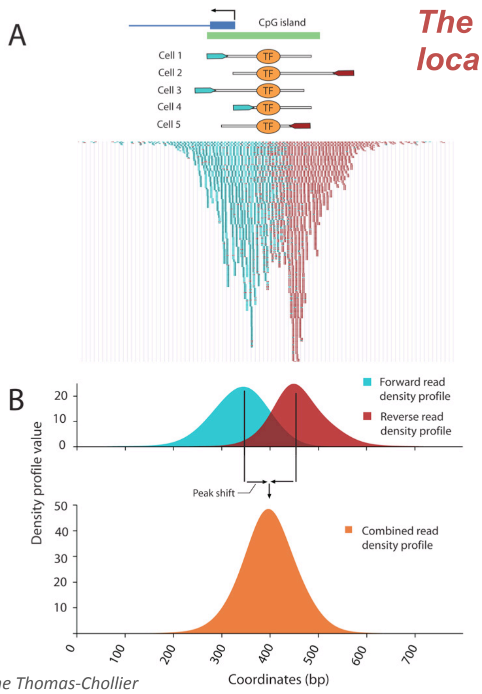
A



The read « peaks » are not the location of the binding site !

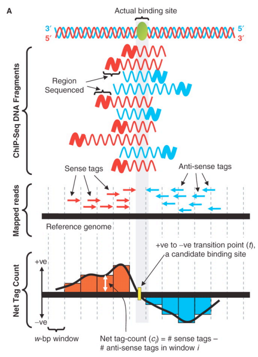
Morgane Thomas-Chollier

Valouev Nat Methods (2008), Jothi, NAR (2008)



The read « peaks » are not the location of the binding site !

mapping



peak-calling

Morgane Thomas-Chollier

Valouev Nat Methods (2008), Jothi, NAR (2008)

Peak-calling step

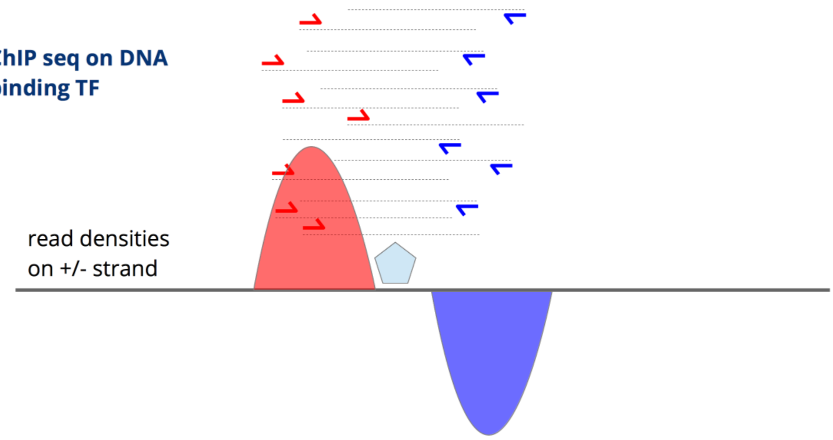
- Treating the reads (tag shifting or elongation)
- Modelling noise levels (input)
- Scaling datasets
- Detecting enriched/peak regions

Morgane Thomas-Chollier

How to determine the position of the TF ?

ChIP seq on DNA binding TF

read densities
on +/- strand



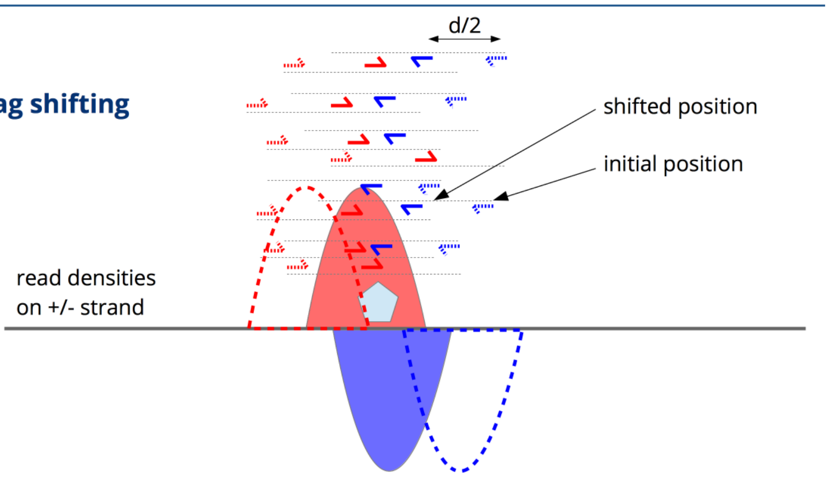
We expect to see a typical strand asymetry in read densities
→ ChIP peak recognition pattern

Carl Herrmann

From aligned reads to binding sites

Tag shifting

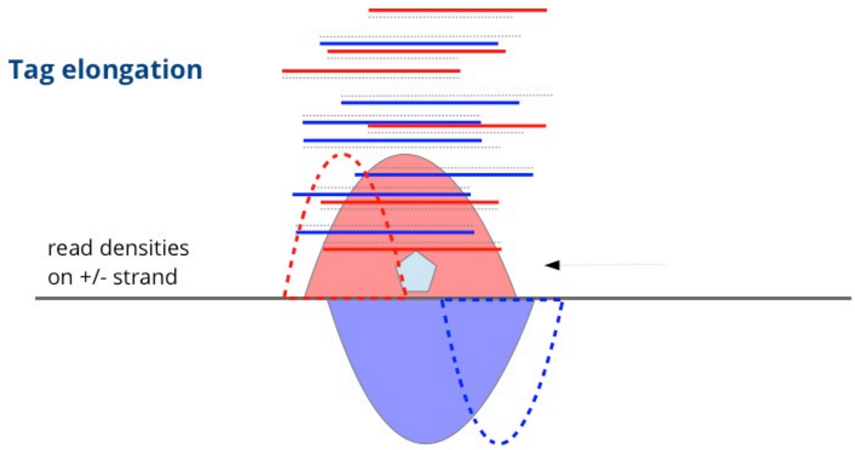
read densities
on +/- strand



Each tag is shifted by *d/2* (i.e. towards the middle of the IP fragment)
where *d* represent the fragment length

Carl Herrmann

From aligned reads to binding sites



Each tag is computationaly extended in 3' to a total length of *d*

Carl Herrmann

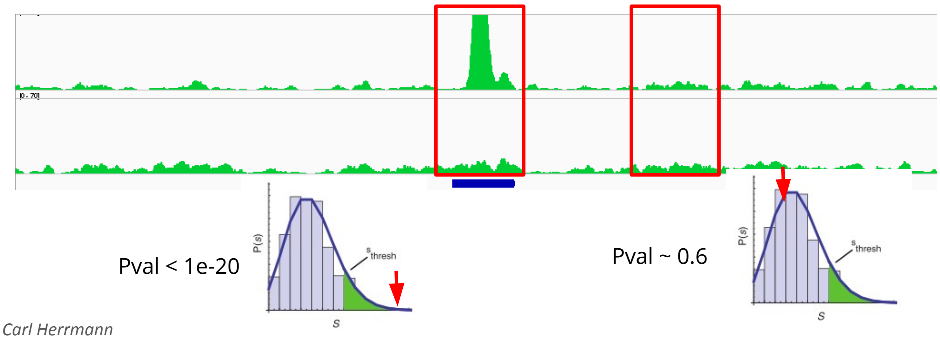
Peak-calling step

- Treating the reads (tag shifting or elongation)
- Modelling noise levels (input)
- Scaling datasets
- Detecting enriched/peak regions

Carl Herrmann

Defining "peaks"

- **Determining "enriched" regions**
  - sliding window across the genome
  - at each location, evaluate the enrichement of the signal wrt. expected background based on the distribution
  - retain regions with P-values below threshold
  - evaluate FDR



Carl Herrmann

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based duplicate[e]
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\# \text{control}}{\# \text{ChIP}}$	Optional peak height, ratio to background	Yes / No
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\# \text{control}}{\# \text{ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\# \text{control}}{\# \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\# \text{control}}{\# \text{ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value	No / Yes
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest pairec tag distance					
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation					

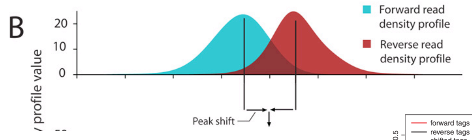
Computation for ChIP-seq and RNA-seq studies

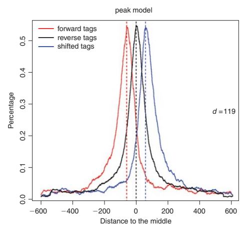
Shirley Pepke[1], Barbara Wold[2] & Ali Mortazavi[2]

Carl Herrmann

Peak-calling with MACS: overview

bimodal enrichment pattern





Two steps strategy :

1 – modelling the read shift size

2 – peak calling

1 : search high-quality paired peaks : separates their forward and reverse reads, and aligns them by the midpoint. The distance between the modes of the forward and reverse peaks in the alignment is defined as *d*, and MACS shifts all reads by *d*/2 toward the 3′ ends to better locate the precise binding sites.

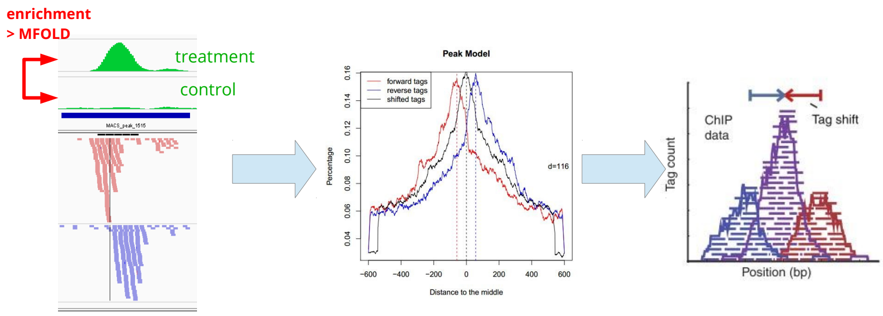
2: uses the shift size to search for peaks, Poisson distribution to measure the p-value of each peak, and False Discovery Rate (FDR) calculation using the input data

Feng, J., Liu, T., & Zhang, Y. (2011). *Using MACS to Identify Peaks from ChIP-Seq Data, Current Protocols in Bioinformatics*

1 – modelling the read shift size **MACS**

[Zhang et al. Genome Biol. 2008]

- **Step 1 : estimating fragment length *d***
  - slide a window of size BANDWIDTH
  - retain top regions with MFOLD enrichment of treatment vs. input
  - plot average +/- strand read densities → estimate d



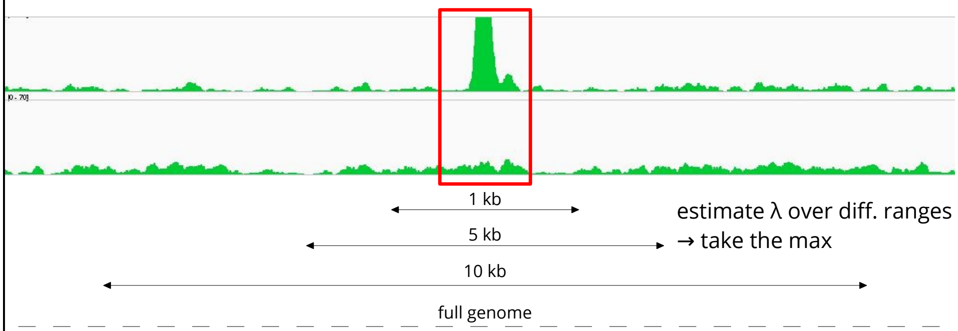
Carl Herrmann

2 – peak-calling

MACS

[Zhang et al. Genome Biol. 2008]

- **Step 2 : identification of local noise parameter**
  - slide a window of size *2*d* across treatment and input
  - estimate parameter λ_{local} of Poisson distribution



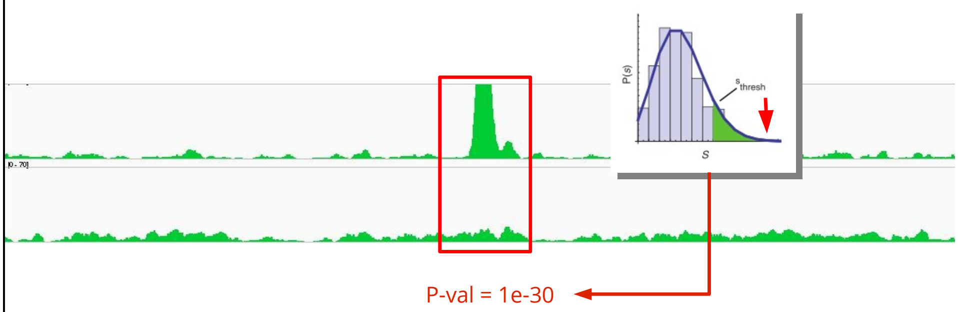
Carl Herrmann

2 – peak-calling

MACS

[Zhang et al. Genome Biol. 2008]

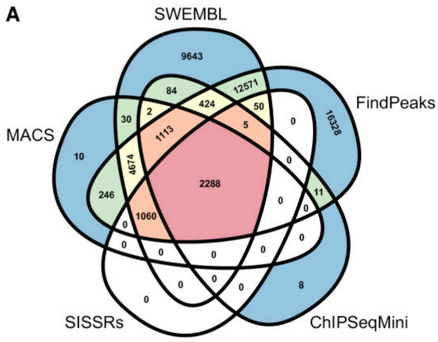
- **Step 3 : identification of enriched/peak regions**
  - determine regions with P-values < PVALUE
  - determine summit position inside enriched regions as max density



Carl Herrmann

Peak-calling programs

- Strong influence on the called peaks
  - Many different programs
  - They do not share the same « default » threshold to retain peaks
  - The top highest peaks are usually common, but the less obvious peaks are often not shared between different peak callers



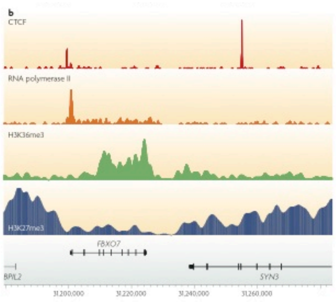
Mali Salmon-Divon *et al*, *BMC Bioinformatics, 2010*

Morgane Thomas-Chollier

Peak-calling programs

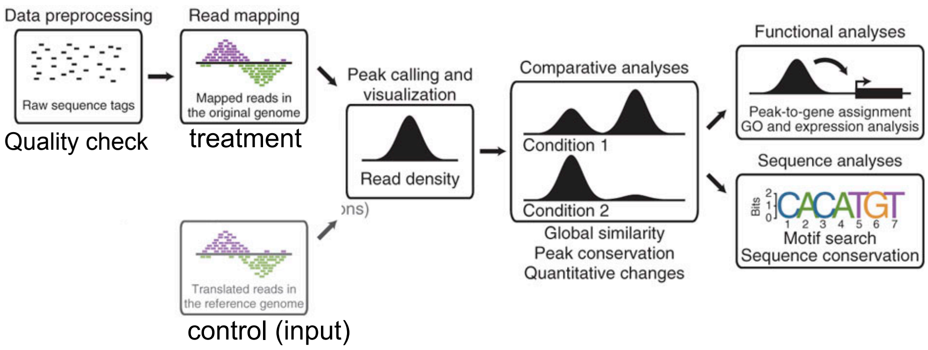
- To be chosen according to type of expected peaks
  - Transcription factors and « sharp » peaks: MACS2 for TF: --call-summits
  - Chromatin marks and « broad peaks » MACS2 --broad

- Many new programs still developped !



Morgane Thomas-Chollier

ChIP-seq analysis workflow



Processing steps

Downstream analyses

Morgane Thomas-Chollier

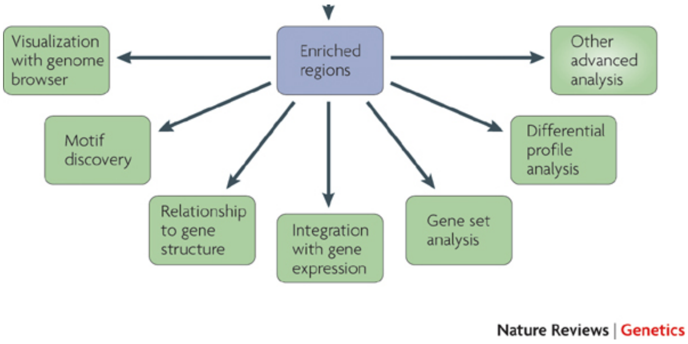
Adapted from Bardet et al, Nature Protocols, 2012

Hands on !

- Go to the companion website
- Follow all steps of **Visualizing the peaks in a genome browser**
- If you have the time, do the **bonus** exercise

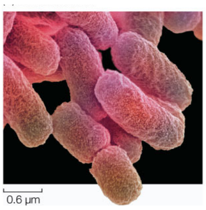
Morgane Thomas-Chollier

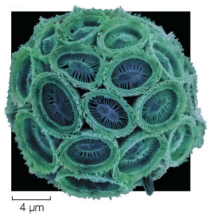
ChIP-seq analysis workflow: downstream analyses



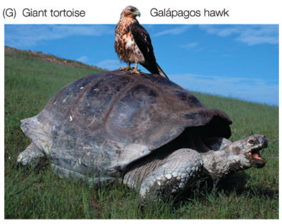
Park, Nature reviews 2009

Morgane Thomas-Chollier





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

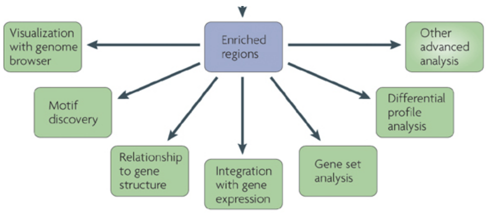
What is the biological question ?

- **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



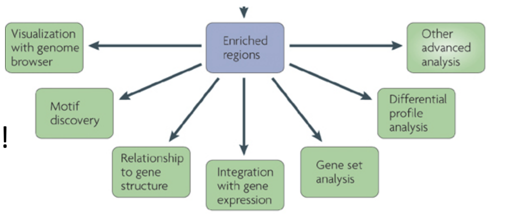
Nature Reviews | Genetics

Morgane Thomas-Chollier

What is the biological question ?

➔ Should drive all « downstream » analyses

Will take time to « do it all » !!!



Nature Reviews | Genetics

Morgane Thomas-Chollier

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

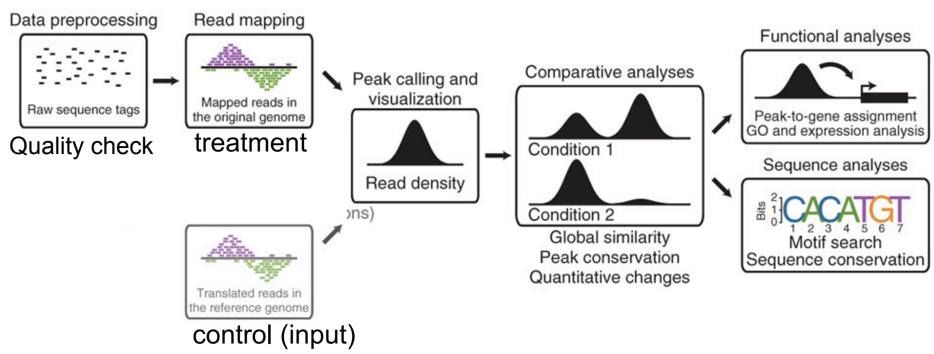
Morgane Thomas-Chollier

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

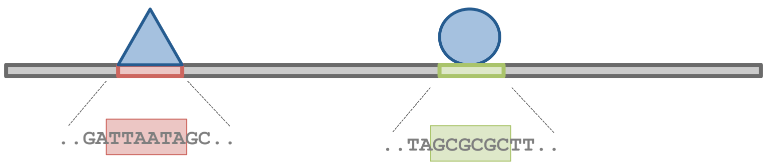


Morgane Thomas-Chollier

Adapted from Bardet et al, Nature Protocols, 2012

Transcription factor specificity

How do TF « know » where to bind DNA ?

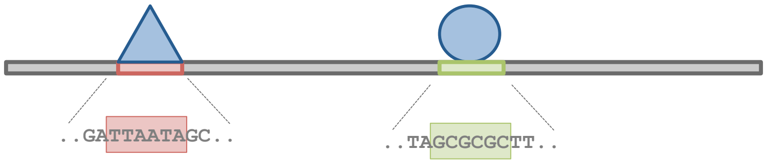


TF recognize TFBS with specific DNA sequences

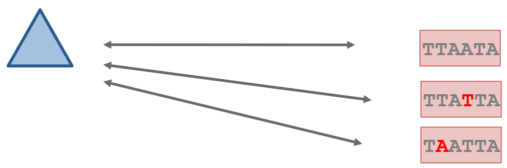
Morgane Thomas-Chollier

Transcription factor specificity

How do TF « know » where to bind DNA ?



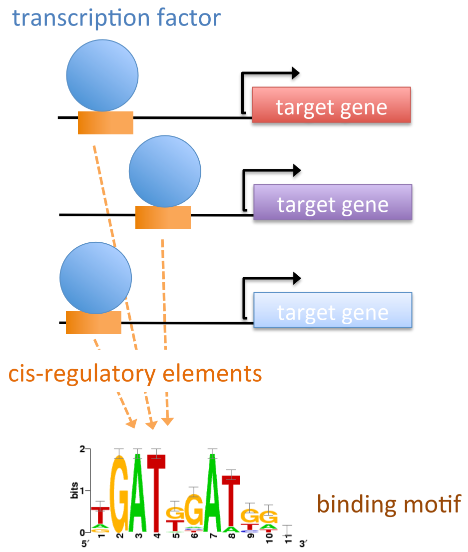
TF recognize TFBS with specific DNA sequences



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

Morgane Thomas-Chollier

de novo motif discovery



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

Morgane Thomas-Chollier

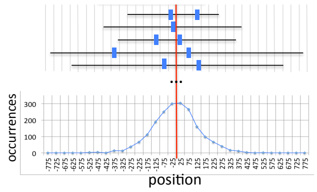
de novo motif discovery

- 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

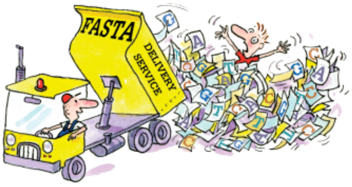
- Tools already exist for a long time !

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

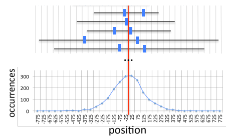
Why do we need new approaches for genome-wide datasets ?

Morgane Thomas-Chollier

New approaches for ChIP-seq datasets



- **Size, size, size**
  - limited numbers of promoters and enhancers
  - dozens of thousands of peaks !!!!!!

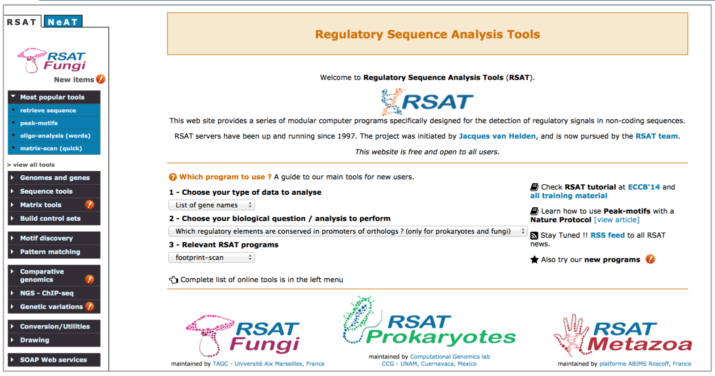


- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias
- **motif analysis: not just for specialists anymore !**
  - complete user-friendly workflows

Morgane Thomas-Chollier

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

RSAT



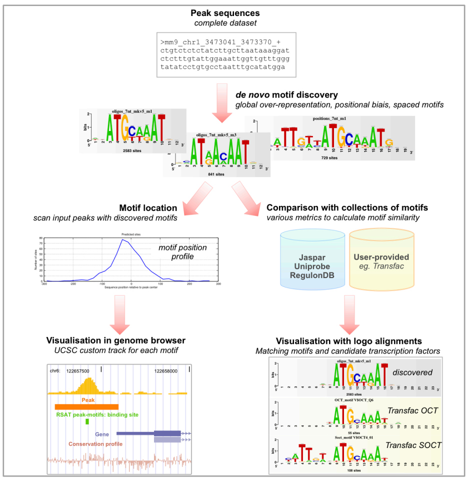
http://rsat.eu

Medina, Defrance, Sand et al Nucleic Acids Research, 2015
Thomas-Chollier et al Nucleic Acids Research, 2011
Thomas-Chollier, Sand et al, Nucleic Acids Research, 2008
van Helden, Nucleic Acids Research, 2003

Morgane Thomas-Chollier

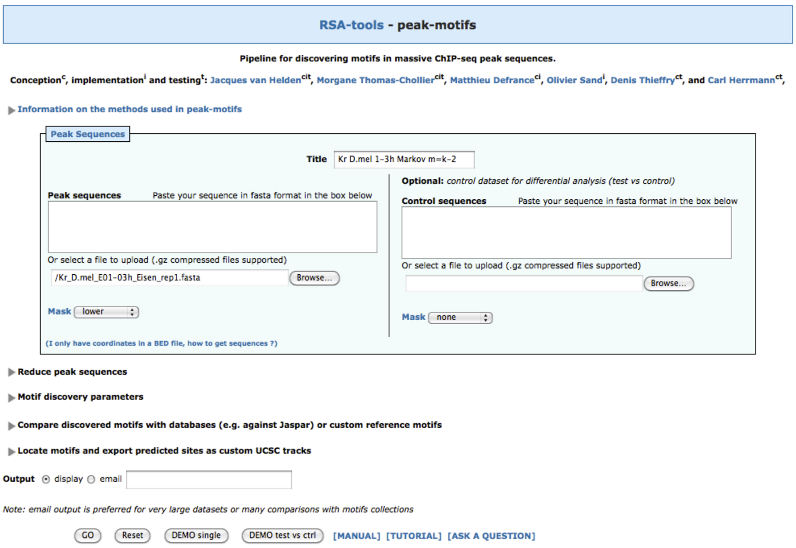
Peak-motifs

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



Thomas-Chollier et al Nucleic Acids Research, 2012

Morgane Thomas-Chollier



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**

  - Demo buttons
  - Tutorials & Protocols

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

  - HTML report

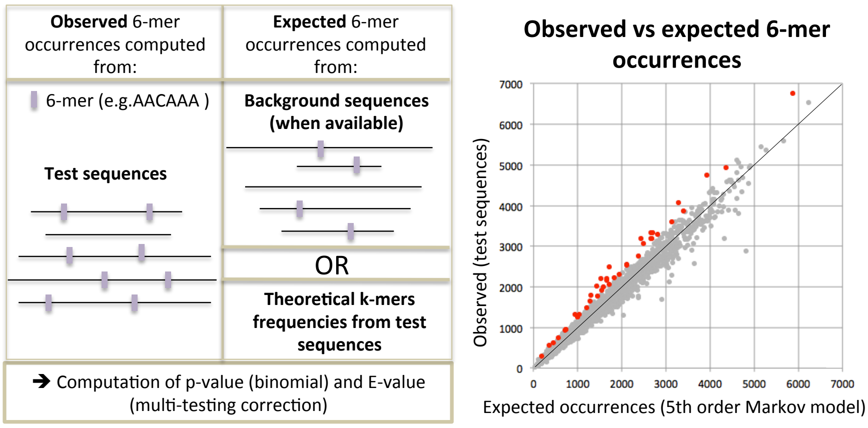
Morgane Thomas-Chollier

Hands on !

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

Morgane Thomas-Chollier

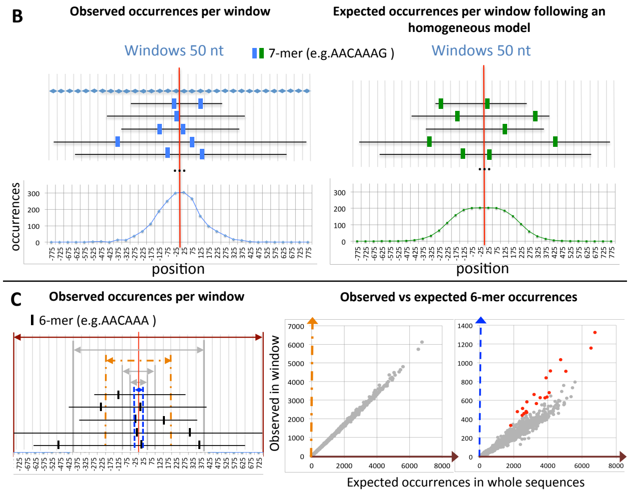
Motif discovery methods: frequency



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

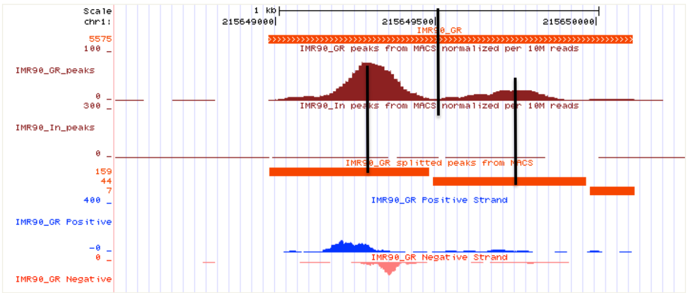


position-analysis

local-words

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

Peaks from MACS



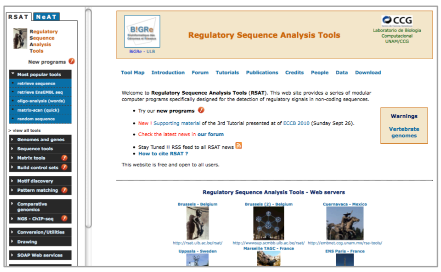
=> Use **peak-splitter** or extract summit +/- 200 bp

Morgane Thomas-Chollier

Acknowledgements

Jacques van Helden
Denis Thieffry
Carl Herrmann
Mathieu Defrance
Olivier Sand
Elodie Darbo

http://rsat.eu



Janick Mathys (VIB) for inviting me for this training !

Possible topics for discussion

It's common practice to sequence the input deeper than the treatment. Why ?

Importance of the mapping tool ?

Single-end or paired-end sequencing ?

ChIP-seq or ChIP-exo ?

Why do we find peaks that do not have two opposite read densities ?

I see ChIP-seq peaks specifically on exons, should I worry ?