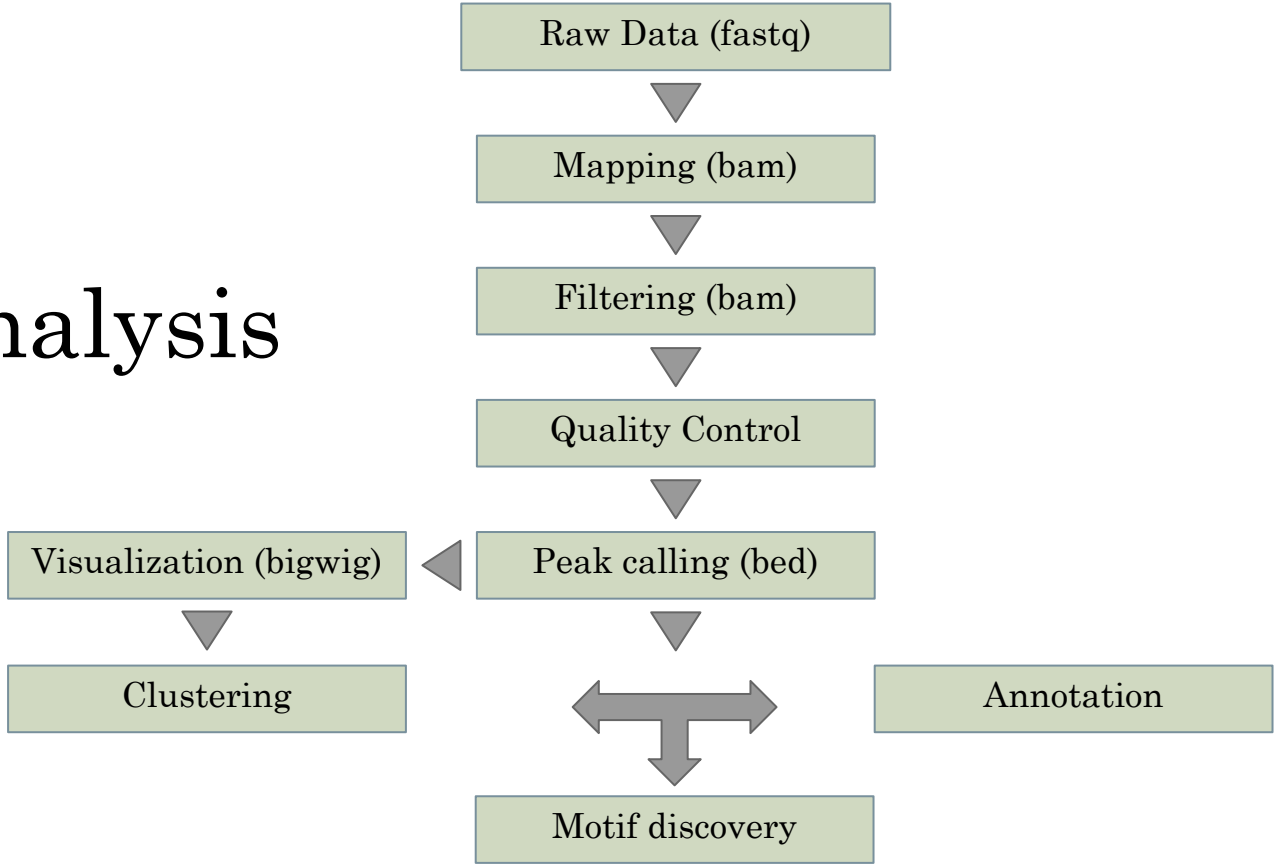
Motif analysis

D. Puthier, C. Rioualen, J. van Helden

Galaxy Workshop — Cuernavaca, 2017

Slides prepared in collaboration with:
Morgane Thomas-Chollier, Carl Herrmann,
Matthieu Defrance, Stéphanie Le Gras

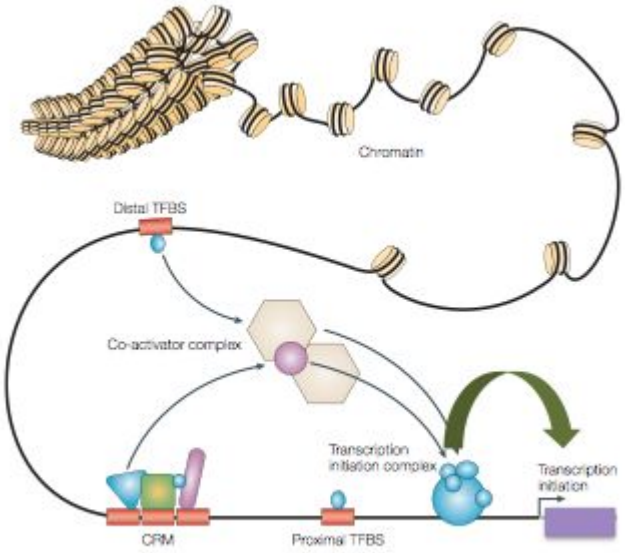
Motif analysis



Discovering motifs in peaks

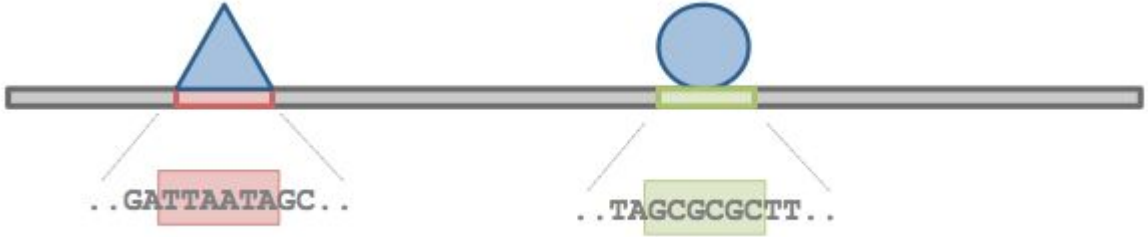
Biological concepts of transcriptional regulation

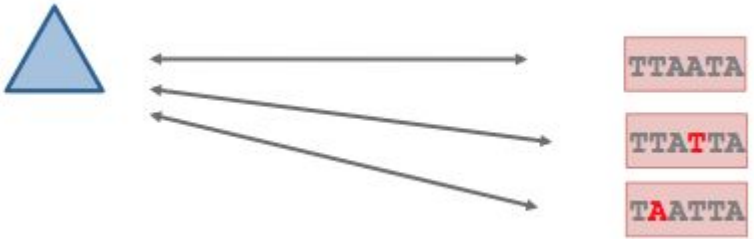
Transcription factors are proteins that modulate (activate/repress) the expression of **target genes** through the binding on **DNA cis-regulatory elements**

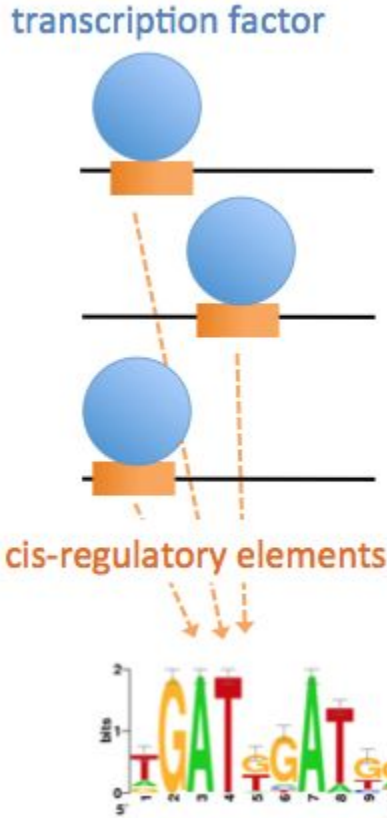


Wasserman et al, Nat Rev Genet, 2004

Transcription factor specificity

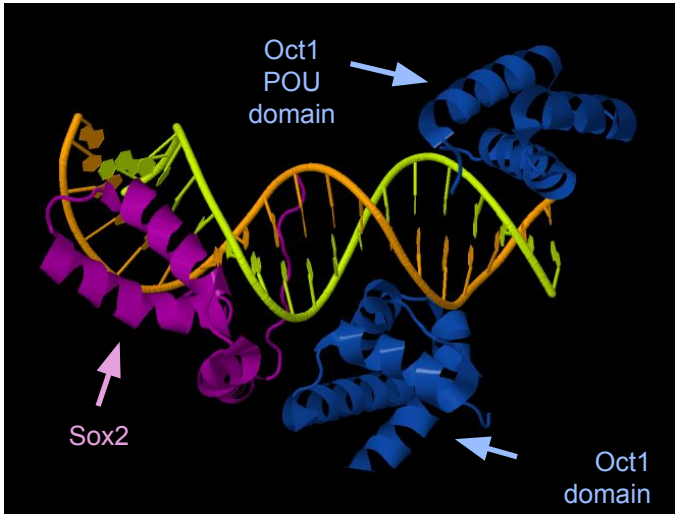






Sox2/Oct4 cooperative binding

- The Sox2 and Oct4 transcription factors recognize specific DNA motifs.
- Cooperative binding: Sox2 and Oct4 closely interact to bind DNA.
- The pair of transcription factors recognizes a composite motif called the « SOCT » motif (SOx+OCT).



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

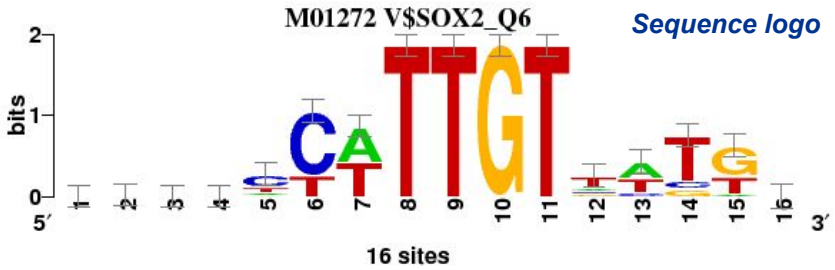
Sox2 : from binding sites to binding motif

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

```
R15133   GCCCTCATTGTTATGC
R15201   AAACTCTTTGTTTGGA
R15231   TTCACCATTGTTCTAG
R15267   GACTCTATTGTCTCTG
R16367   GATATCTTTGTTTCTT
R17099   TGCACCTTTGTTATGC
R19276   AATTCCATTGTTATGA
R19367   AAACTCTTTGTTTGGA
R19510   ATGGACATTGTAATGC
R22342   AGGCCTTTTGTCCTGG
R22344   TGTGCTTTTGTNNNNN
R22359   CTCAACTTTGTAATTT
R22961   GCAGCCATTGTGATGC
R23679   CACCCCTTTGTTATGC
R25928   TTTTCTATTGTTTTTA
R27428   AAAGGCATTGTGTTTC
```

Position-specific scoring matrix (PSSM)

A	6	7	4	4	2	0	**8**	0	0	0	0	2	**7**	0	1	4
C	2	2	6	5	**9**	**12**	0	0	0	0	0	2	2	2	0	6
G	4	3	2	4	1	0	0	0	0	**16**	0	2	0	2	**9**	3
T	4	4	4	3	**4**	**4**	**8**	**16**	**16**	0	**16**	**9**	**6**	**11**	**5**	2



“Family” binding motifs (FBM)

- In addition to TF-specific matrices, TRANSFAC contains matrices representing the “consensus” of the binding specificity for several transcription factors belonging to the OCT family.
- This matrix was built from 55 sites, collected from different organisms (mouse, human, cat, xenopus, ...).

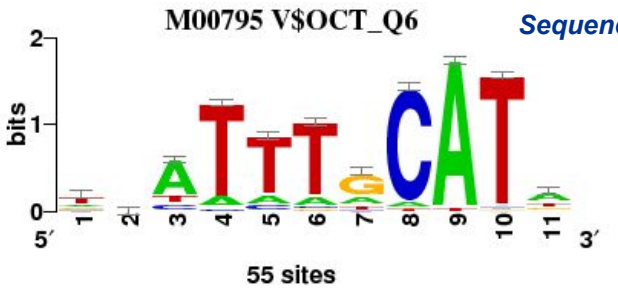
Collection of binding sites
used to build the motif of the OCT ***family*** *(TRANSFAC M00795)*

```
R00306TAATTAGCATA
R00551ATATTTGCATT
R00662TTATTTGCATA
R00664TCATTTGCATA
R00666ACATTTGCATA
R00814TCGTTAGCATG
R00815CGCATGGCATC
R00820GGAATTCCATT
R00824CGTATCTCATT
R00834TTATTTGCATA
R00842GGATTTGCATA
R00855GTATTTGCATA
R00872TAATTTGCATT
R00888CGATTTGCATA
R00893TGATTTGCATA
... 40 other sites
```

Position-specific scoring matrix (PSSM)

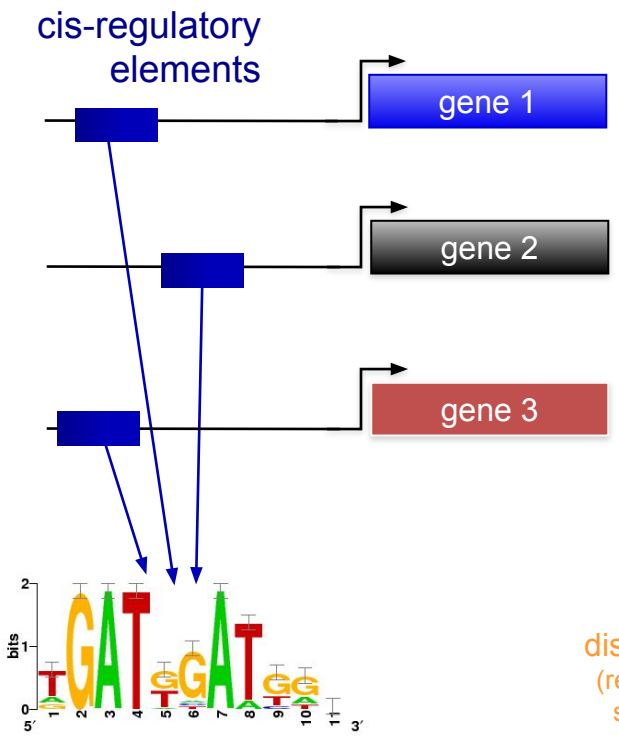
A	10	14	**37**	6	7	6	**11**	3	**53**	1	**27**
C	7	12	7	2	5	2	3	**50**	0	1	4
G	10	15	2	0	1	2	**34**	0	0	1	10
T	**28**	14	9	**47**	**42**	**45**	7	2	2	**52**	14

Sequence logo

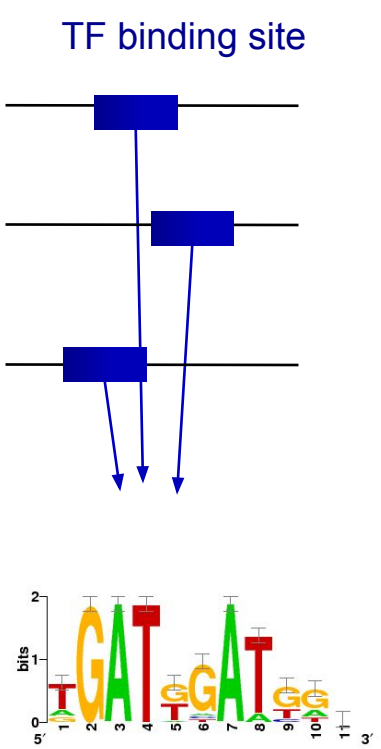


De novo motif discovery

Case 1: promoters of co-expressed genes



Case 2: ChIP-seq peaks



discovered motif
(represented as a sequence logo)

De novo motif discovery

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

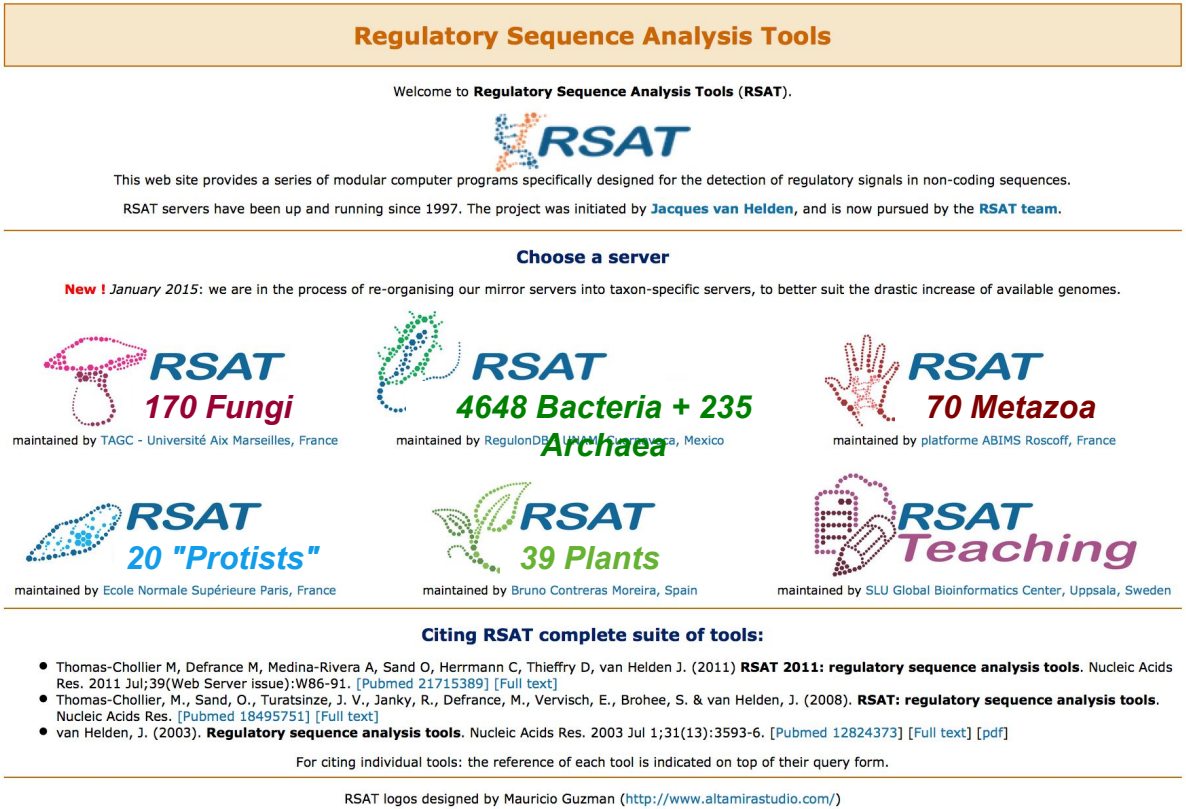
Some motif discovery tools

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

Motif analysis on ChIP-seq peaks

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

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



Contributors From ULB

B!GRe
Bioinformatique des Génomes et Réseaux


Sylvain Brohée
Postdoc


Nicolas Simonis
Postdoc


Didier Croes
Postddoc


Didier Gonze
Premier assistant

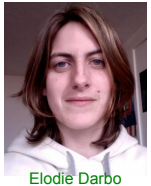

Myriam Loubriat
Secretary


Ariane Toussaint
Professor Emeritus


Jacques van Helden
Professor


Leon Juvenal HajingaboE
PhD Student


Maud Vidick
PhD Student (co-direction)


Elodie Darbo
PhD Student
co-direction Marseille


Alejandra Medina
PhD Student
co-direction Mexico


Morgane Thomas-Chollier
PhD student+postdoc


Matthieu Defrance
Postdoc


Olivier Sand
Postdoc

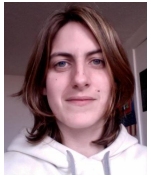

Jean Valéry Turatsinze
PhD student

Collaborators

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

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

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



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

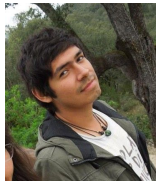
Julio Collado-Vides
(CCG, Cuernavaca - Mexico)
Initiation of the RSAT project
Analysis [illegible]n in bacter[illegible]s

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

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

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

Bruno Contreras
(CSIC, Saragossa, Spain)

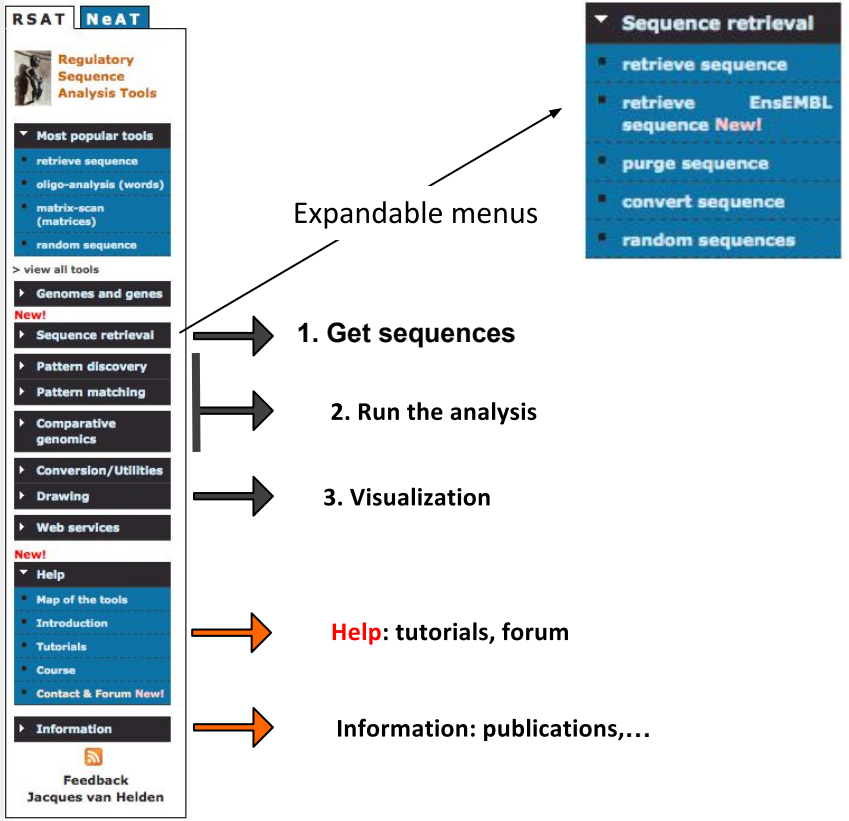


Jaime Castro-Mondragon
(PhD at TAGC, Marseille, France)

Protocol

RSAT Quick Tour

1. Open a connection to the RSAT portal (http://rsat.eu/).
2. Select the **Metazoa server** (hosted on the ABIMS platform at Roscoff - France).
3. Explore the tools
   a. On the home page, use the 3-questions guidance (**Which program to use?**) and find a way to discover motifs in ChIP-seq peaks.
   b. In the left-side menu, click on the black boxes to **expand thematic lists of tools**, and browse the tool names to get a global idea of the supported functionalities.



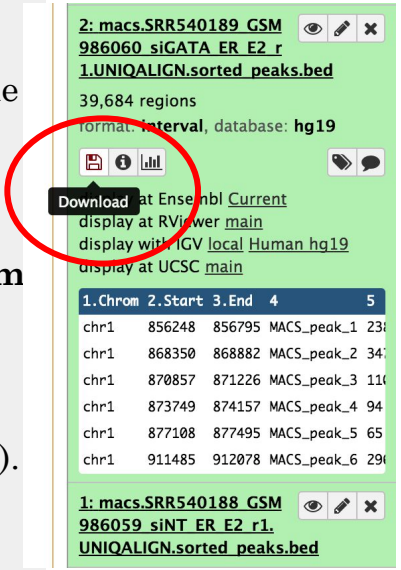
Protocol

Fetch sequences from a bed file

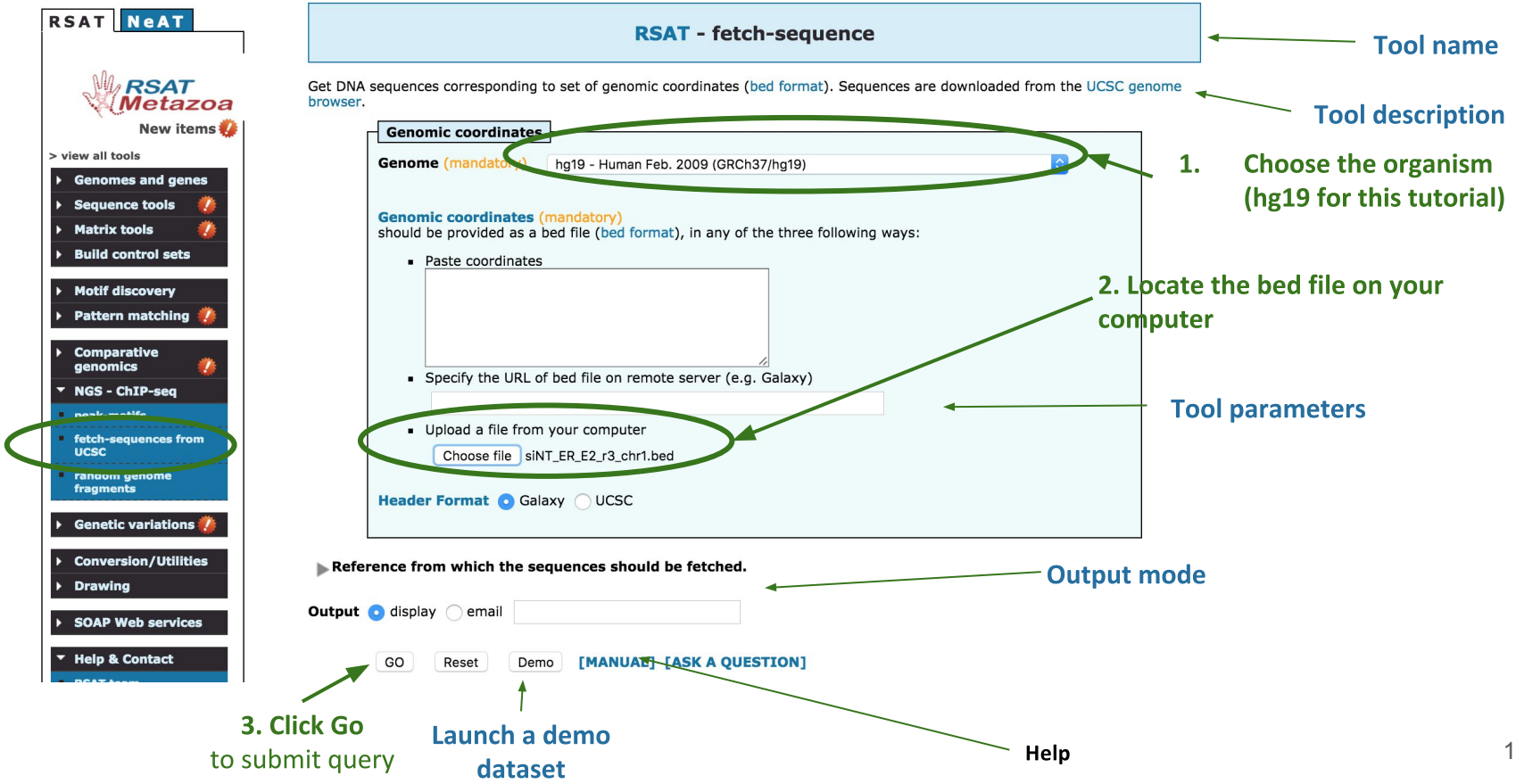
(quick tutorial: see snapshot on next slide)

1. On Galaxy, **download the bed file** produced by MACS for the sample siGATA_ER_E2_r1.
2. In a web browser window open the **RSAT Metazoa server** (http://metazoa.rsat.eu/).
3. In the menu (left side) click **NGS-ChIP-seq** and select **fetch-sequences from UCSC**.
4. Select the genome of interest, in this case: **human, assembly hg19**.
5. Genomic coordinates can be provided in 3 alternative ways:
   - Paste the content of the BED file (not very convenient for large peak sets).
   - Select an URL.
   - Upload a file from your computer **(select your MACS peaks).**
6. Click on the button "**GO**".
7. Click on the link to the log file, and look for the number of sequences retrieved.
8. The FASTA file contains the sequences. For the sake of safety and tractability, download it (right click, save as...) on your computer to keep a copy.

Q: Analyse the result of this fetch-sequences tool (number of sequences, …).

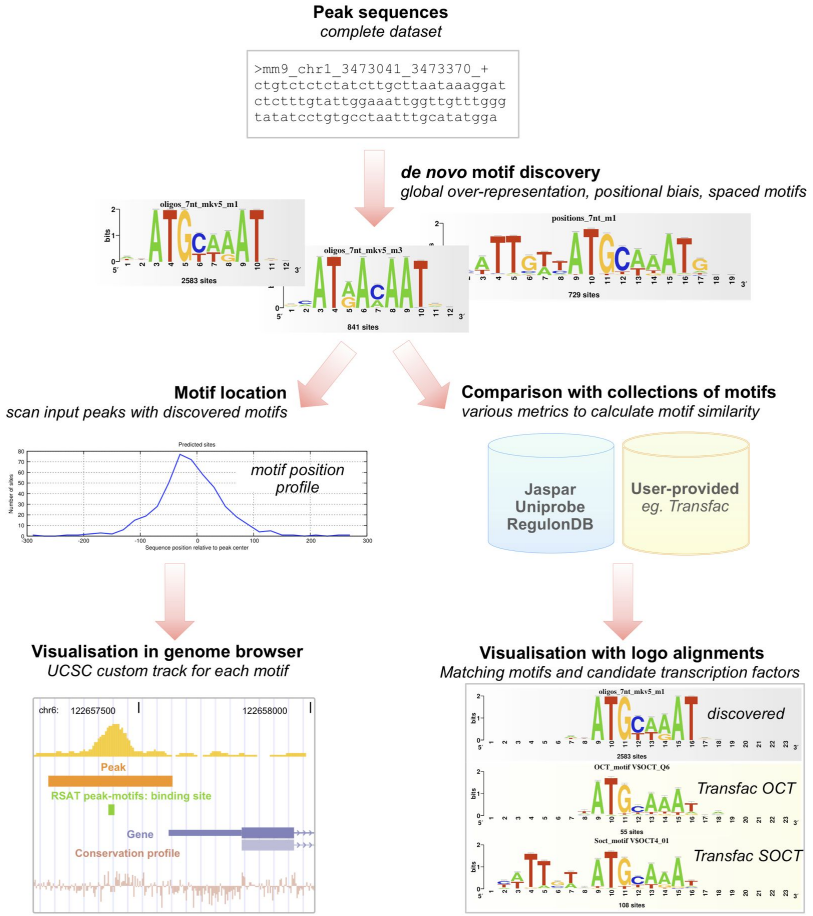


RSAT Web forms - fetch-sequences example



Peak-motifs

- A workflow enabling to discover motifs in large sequence sets (tens of Mb) resulting from ChIP-seq experiments.
- ***Complementary pattern discovery criteria***
  - Global over-representation
  - Positional biases
  - Local over-representation
- Links ***from motifs to putative binding factors***
  - motif databases
  - user-specified reference motifs
- ***Prediction of binding sites*** within the peaks.
  - Inspect distribution around peak centers
  - Can be loaded as UCSC track
- ***Interfaces***
  - Web interface
  - Stand-alone (Unix command-line)
  - Web services (SOAP/WSDL)
  - Virtual Machine for VirtualBox
  - Virtual machine at the IFB cloud
  - *Soon: Debian package*
  - *Soon: Docker container*



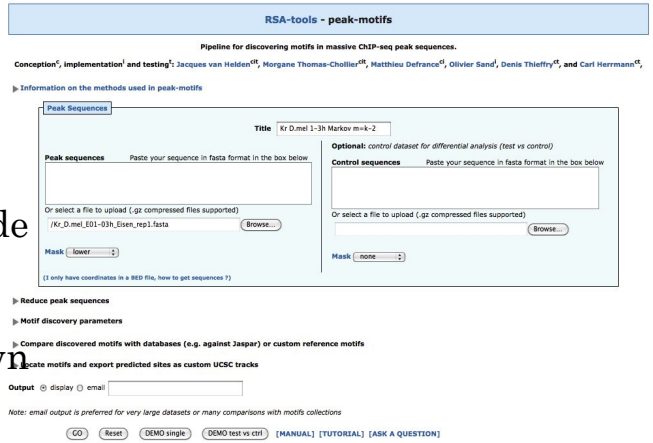
1. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4): e31.
2. Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, **7**, 1551–1568.

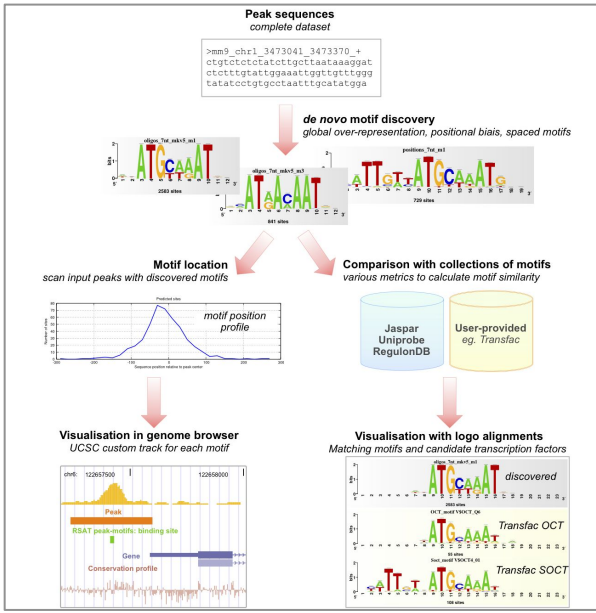
Peak-motifs: why providing yet another tool?

Program	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs	RSAT *peak-motifs*
Web interface	yes	yes	yes	no	no	yes
Size limitation	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	-	unrestricted (Web site tested with 22 Mb)
Stand-alone version	yes	no	yes	yes	yes	yes
Tasks						
peak finding	no	no	no	yes	no	no
annotation of peak-flanking genes	no	yes	no		no	no
sequence composition (mono- and di-nucleotides)	no	no	no		no	yes
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	yes	yes		no	no
enrichment in discovered motifs	no	no	no		no	yes
peak scoring	no	no	yes	yes	no	no
motif clustering	no	no	no		yes	no
comparison discovered motifs / motif DB	no	no	yes		yes	yes
sequence scanning for site prediction	no	no	yes		no	yes
positional distribution of sites inside peaks	no	yes	no		yes	yes
visualization in genome browsers	no	yes	no		no	yes
Motif discovery algorithms	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk

Peak-motifs: why providing yet another tool?

- **Fast and scalable**
- **Treat full-size datasets**
- **Complete pipeline**
  - Peak properties (nucleotide, dinucleotide composition, lengths)
  - Motif discovery
  - Comparison with known motifs
  - Peak scanning
- **Accessible to non-specialists**
  - Demo buttons
  - Tutorials & Protocols
  - Human-readable HTML report with links to all result files.





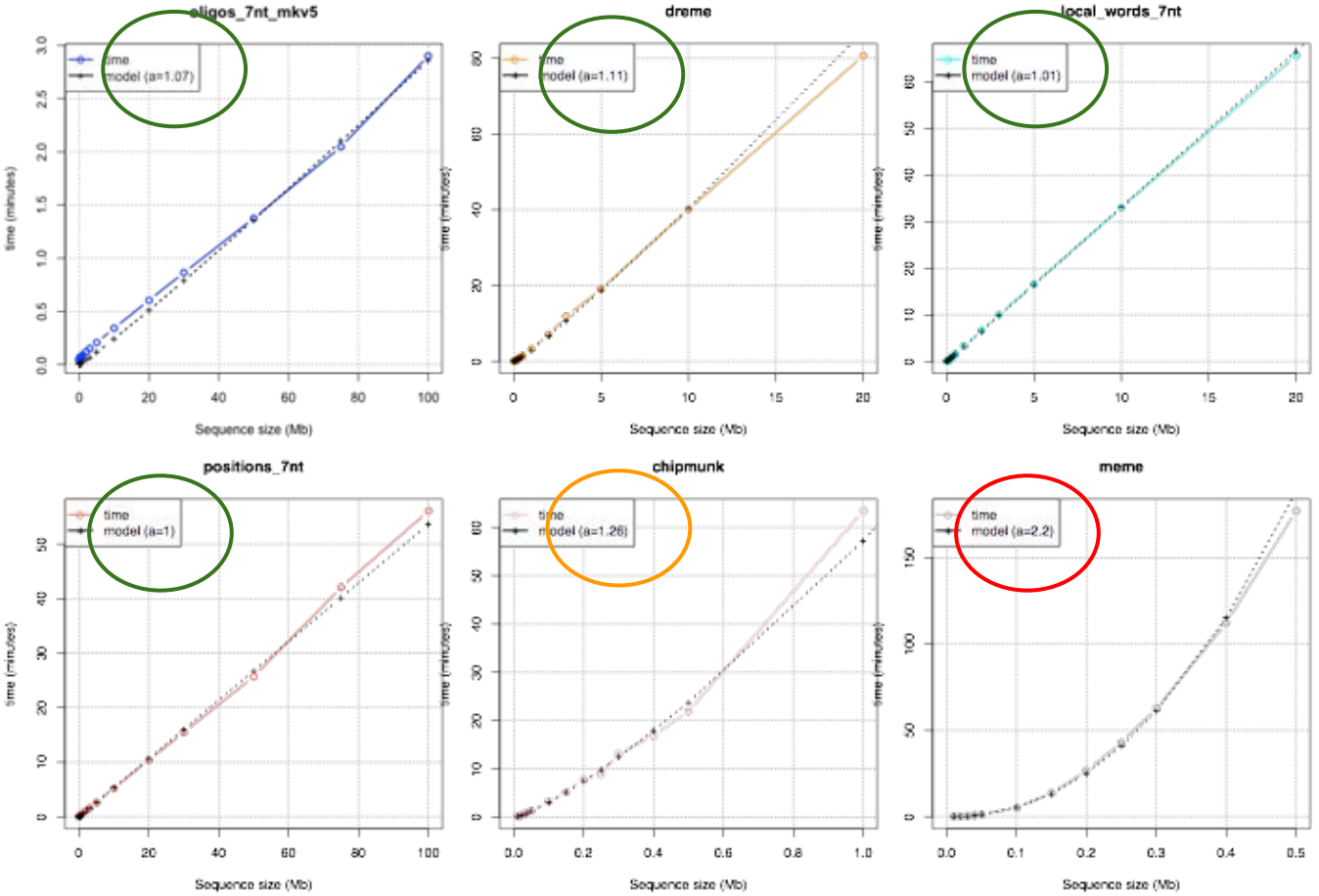
Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden **Nucleic Acids Research**, 2012

Time complexity of motif discovery algorithms

Linear

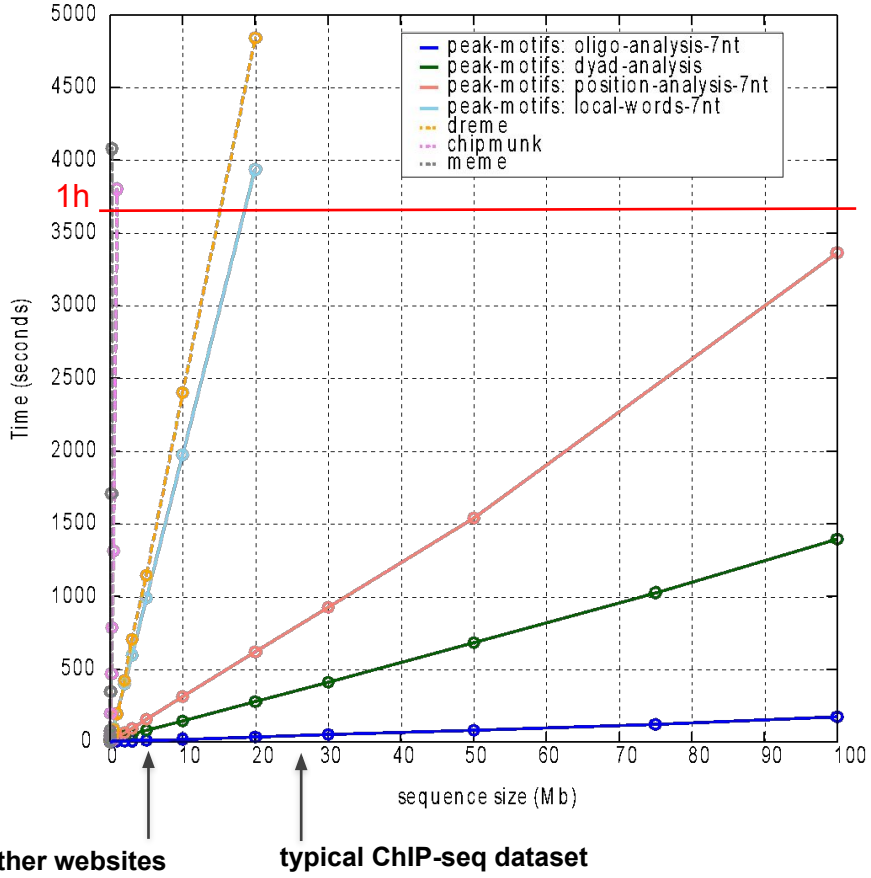
> linear

> quadratic



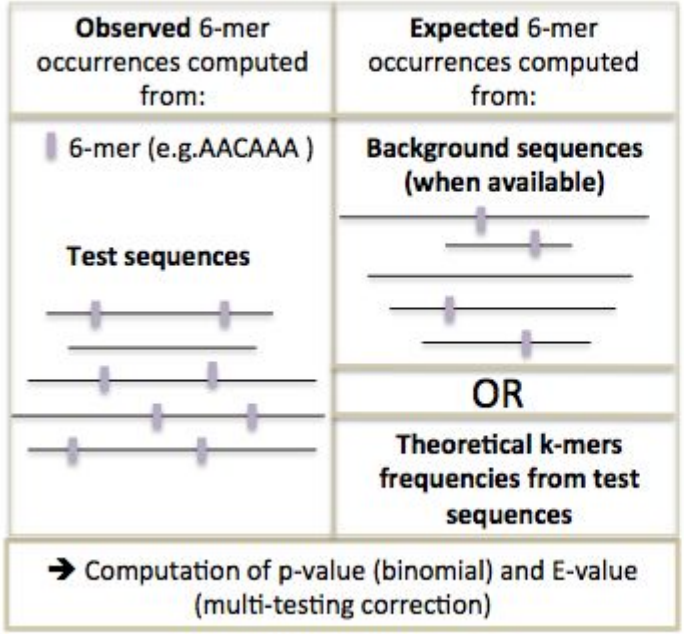
Peak-motifs: scalability

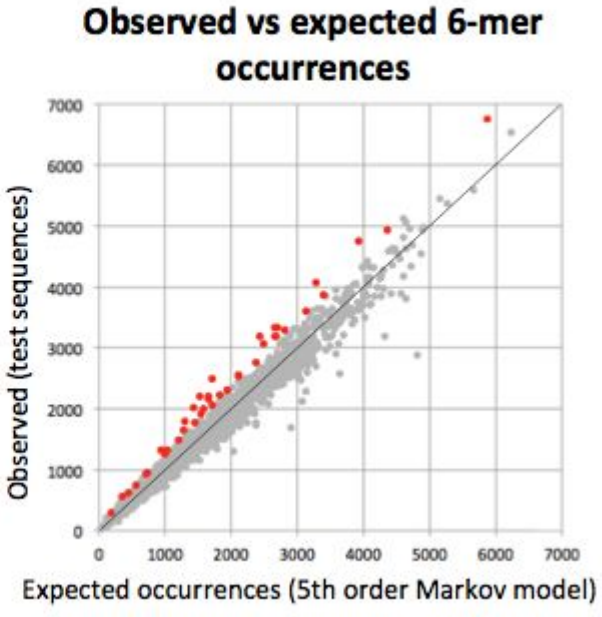
- **Fast and scalable**
- **Treat full-size datasets**
- **Using 4 complementary algorithms**
  - **Global over-representation**
    - **oligo-analysis**
    - **dyad-analysis (spaced motifs)**
  - **Positional bias**
    - **position-analysis**
    - **local-words**



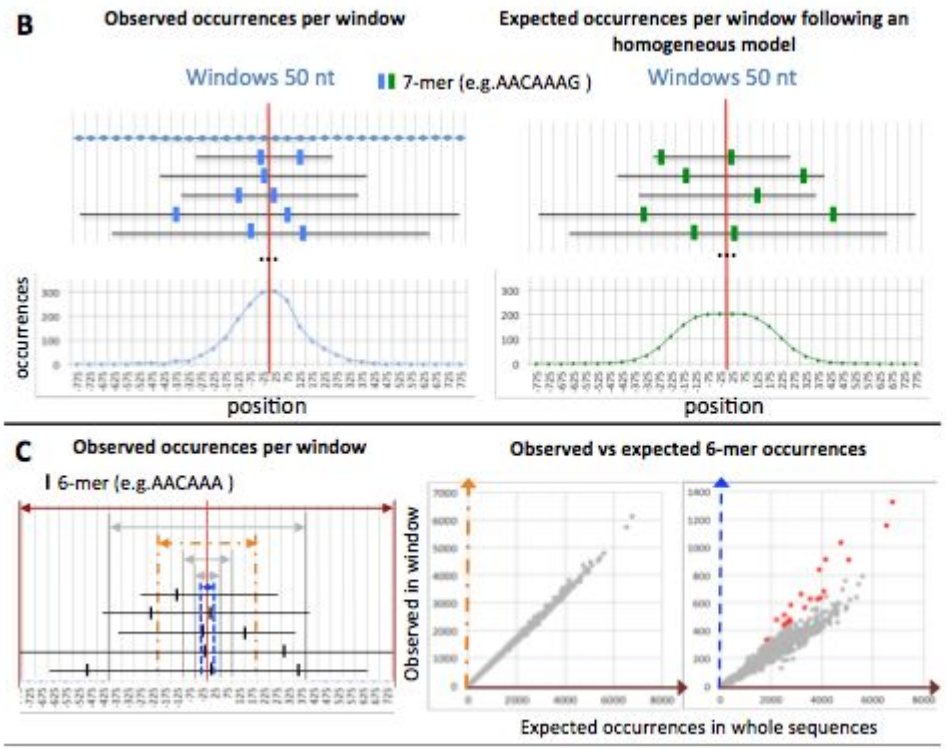
Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden **Nucleic Acids Research**, 2012

Motif discovery: k-mer over-representation





Motif discovery: k-mer position biases

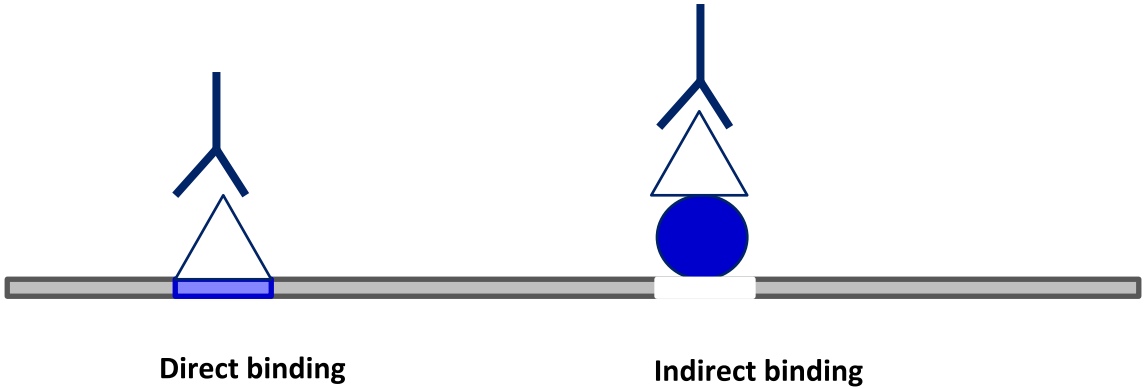


position-analysis

local-words

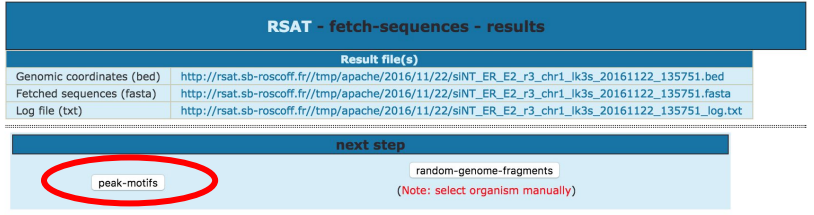
Direct versus indirect binding

- ChIP-seq does not necessarily reveal **direct binding:** The motif of the targeted TF is not always found in peaks!



M2: Getting to know peak-motifs

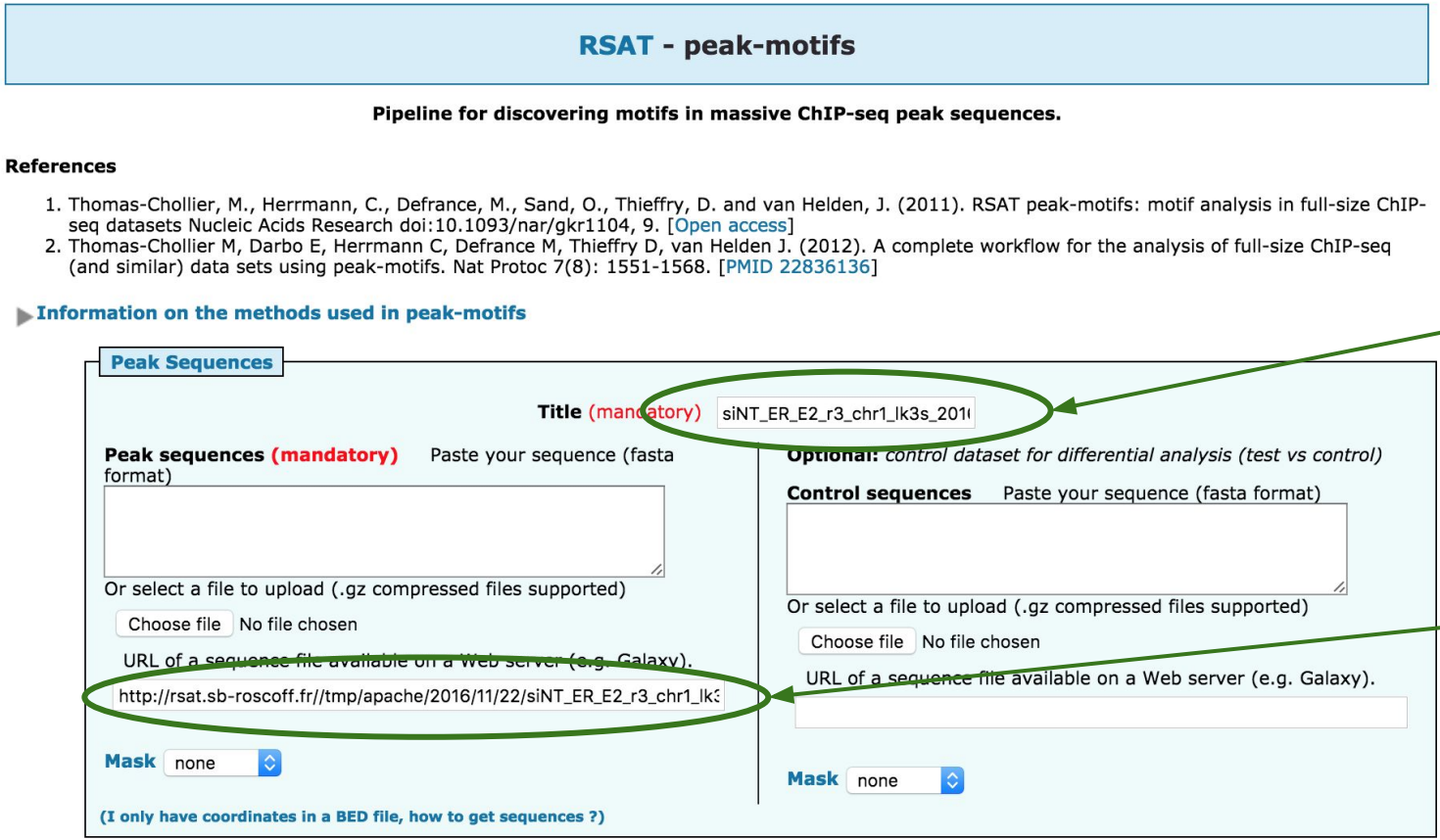
(quick tutorial: see snapshot on next slide)



Protocol

1. In the previous section, we used the RSAT tool **fetch-sequences** to retrieve the peak sequences.
2. At the bottom of the fetch-sequences result page, click on the **peak-motifs** button. A new page appears, displaying a form.
   - *Note: peak-motifs is also accessible from the left menu, in the NGS ChIP-seq, but this would give you an empty form, whereas the Next step button automatically transferred the fetched sequences to the peak-motifs form.*
3. The default peak-motifs web form only displays the essential options, with only two mandatory parameters:
   - **Title box**: type a meaningful name for your peaks (e.g. siGATA_ER_E2_r3)
   - **Sequences:** have been automatically passed from fetch-sequences.
     - *Alternatively, sequences can be pasted in the available box, input from a URL, and uploading a file from your computer.*
4. At this stage you could already possible launch the analysis like (with all other default options), but we will modify some of the **advanced options** in order to fine-tune the analysis according to our data set.

RSAT Web forms: peak-motifs minimal options



Some meaningful title for this analysis

Peak sequences (in fasta format)

URL fed by *fetch-sequences*

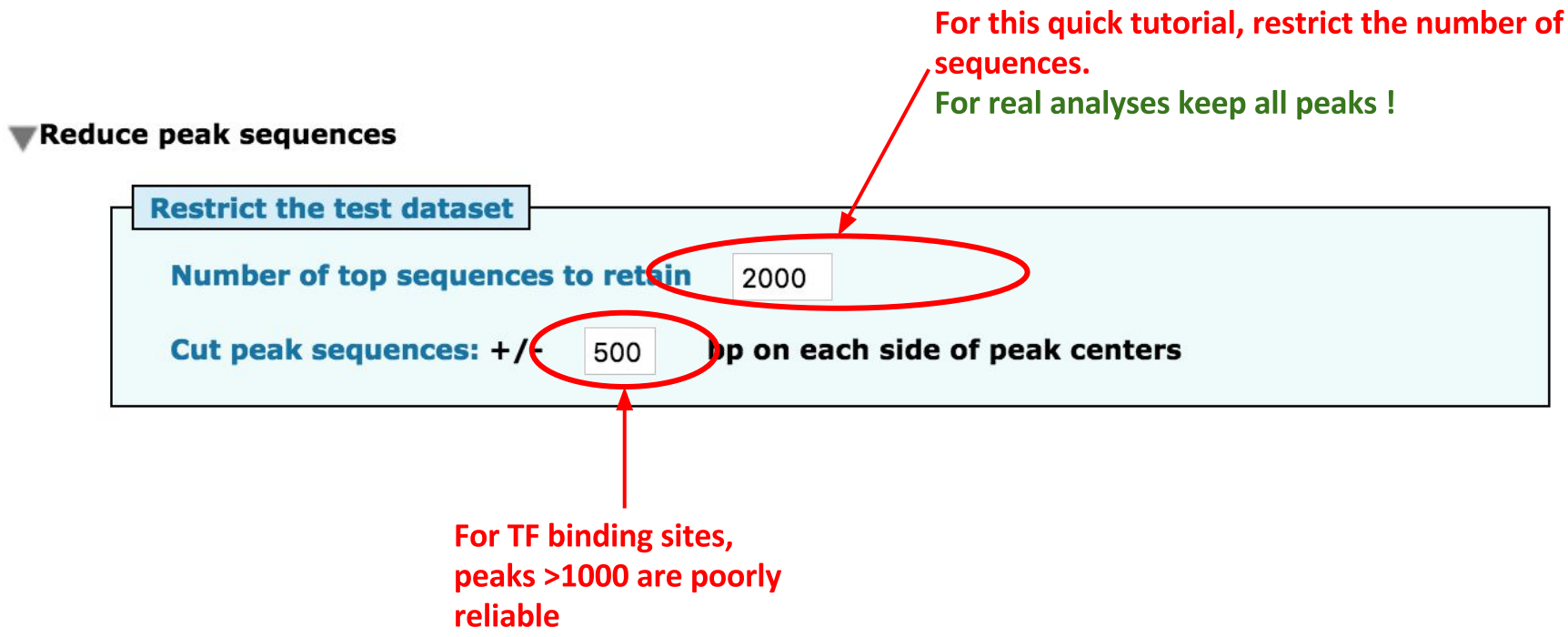
peak-motifs: Launch the analysis

(quick tutorial: see snapshots on next 3 slides)

Protocol

1. Open the ***Reduce peak sequences*** box.
   - Peaks are clipped to 500bp on each side because longer peaks are questionable
   - For a quick tutorial, we suggest to retain the 2000 top peaks only (come back after the course with full datasets).
2. Under ***Motif Discovery parameters***
   - check the **oligomer sizes** 6 and 7 (but not 8).
   - check Discover over-represented spaced word pairs [dyad-analysis]
3. Under ***Compare discovered motifs with databases***,
   - keep ***JASPAR core vertebrates*** as the studied organism is human, and add ***Hocomoco Human***.
4. Under ***Locate motifs and export predicted sites as custom UCSC tracks***
   - select *Peak coordinates in* ***Galaxy/UCSC*** *format (also for* ***fetch-sequences*** *output)*
5. Indicate your ***Email address*** in order to receive notification of the result URL.
   - This is particularly useful because the full analysis may take some time for very large datasets.
6. **Click GO**
   - As soon as the query has been launched, you should receive an email confirming the task submission.
   - **Click on the result link**. The page will be updated from time to time to display intermediate results during the processing.

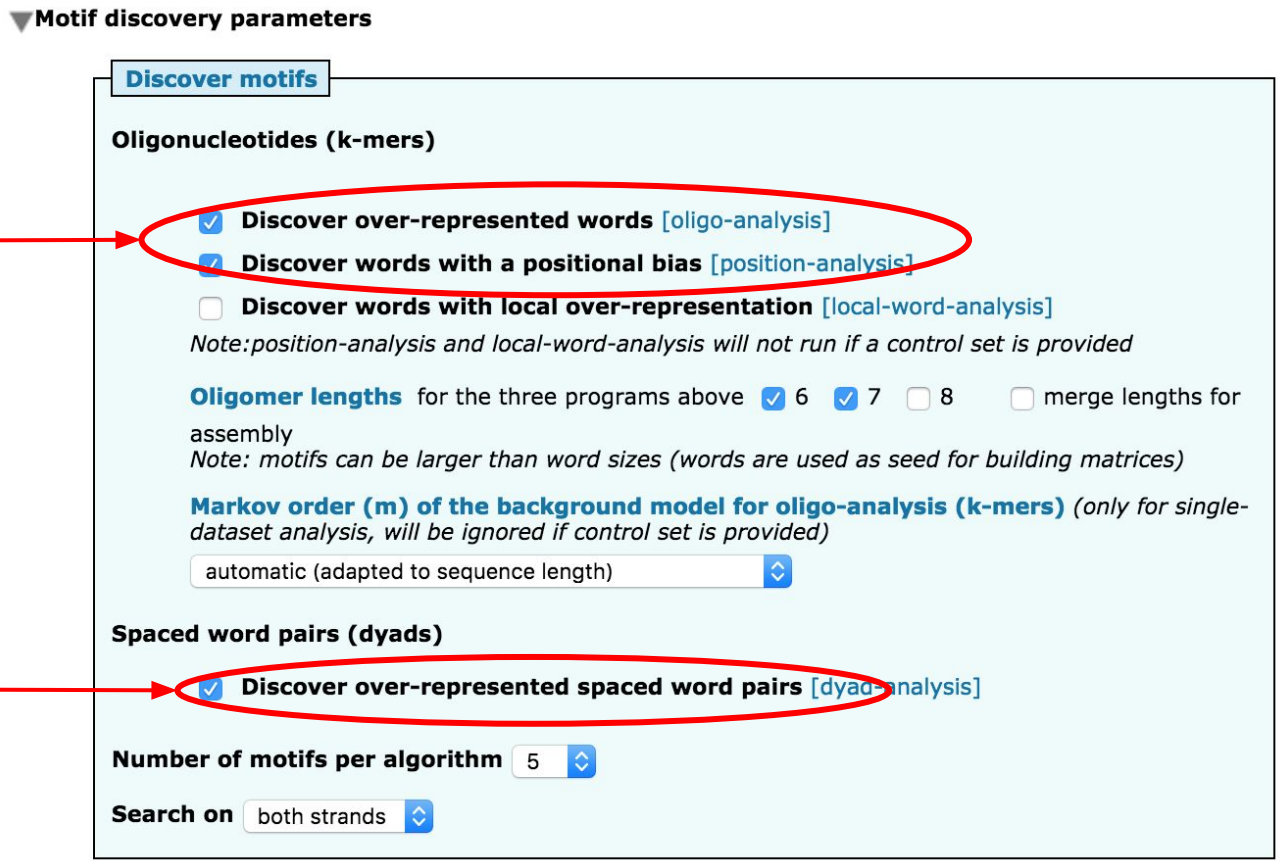
RSAT Web forms: peak-motifs advanced options



RSAT Web forms: peak-motifs advanced options

Two complementary criteria or motif discovery

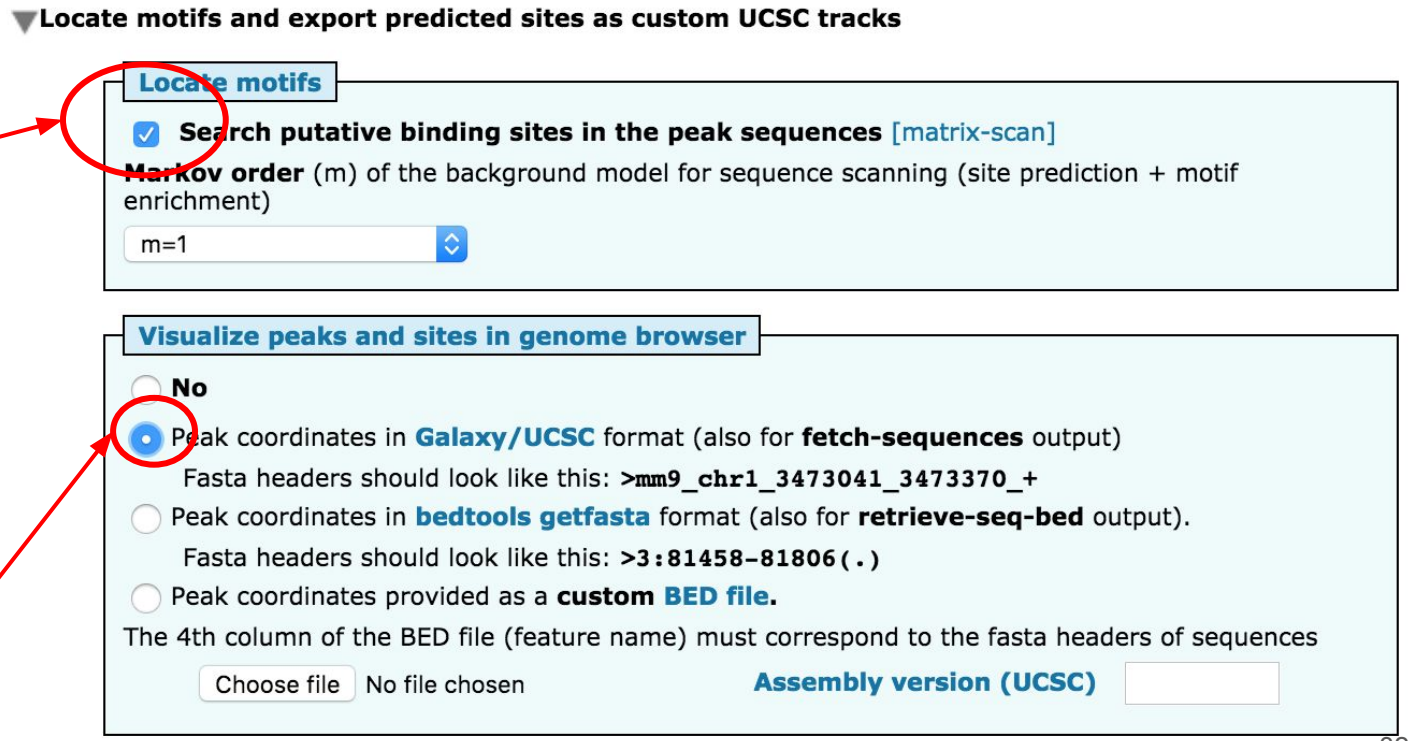
Optionnally, activate dyad-analysis



RSAT Web forms: peak-motifs advanced options

Activate this option to obtain binding site prediction + positional profiles of hits in the peaks.

Enable to upload the sites as annotation tracks on the UCSC genome browser



Negative Controls

Negative Controls in biology

One example from a multitude: Wellik and Mario R Capecchi, Science, 2003.

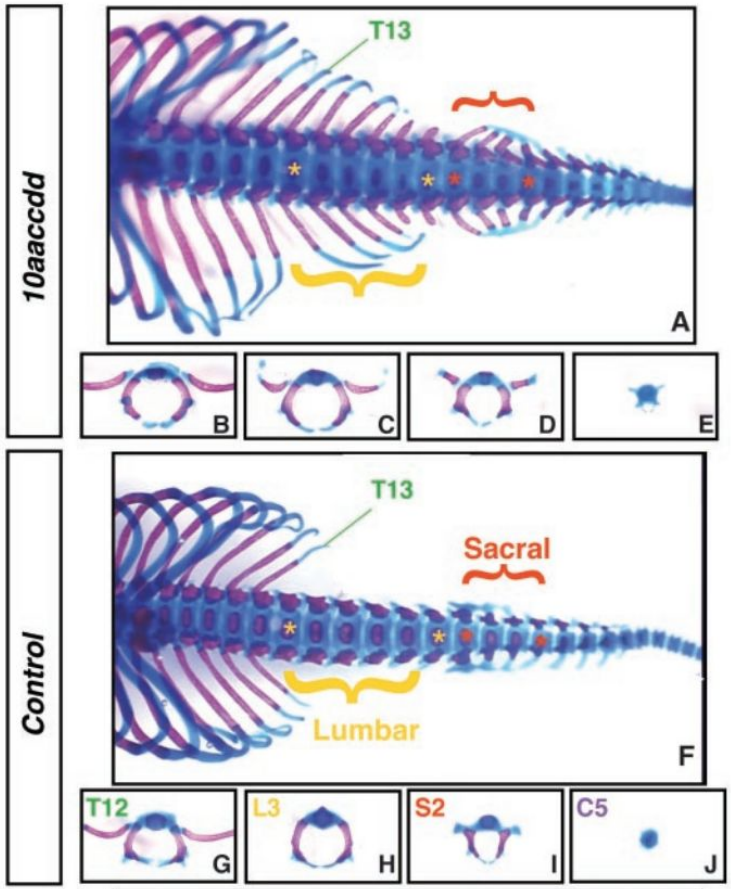
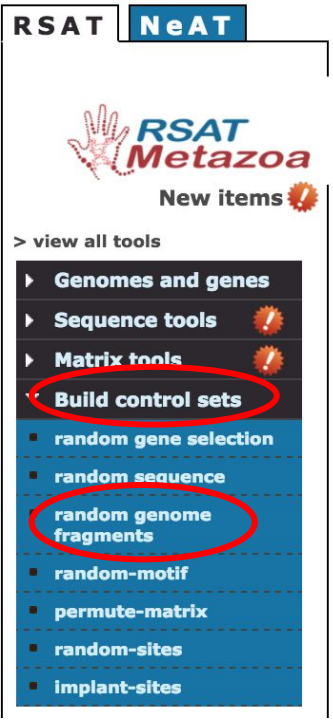
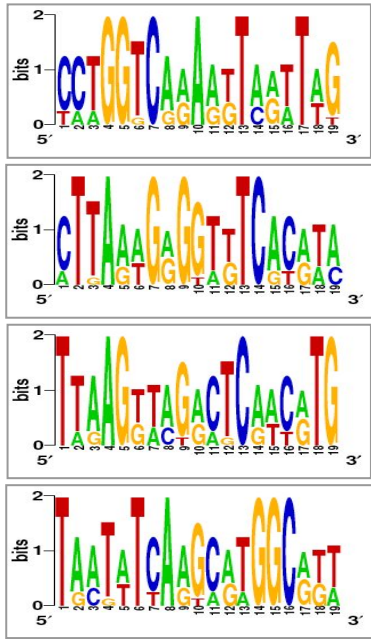


Fig. 1. Axial skeletons of *Hox10* and *Hox11* triple mutants at embryonic day 18.5 (E18.5). Ventral views of the axial skeleton from the lower thoracic region through the early caudal region of a *Hox10* triple mutant (**A**), a control (**F**), and a *Hox11* triple mutant (**K**) are shown. Yellow asterisks indicate lumbar vertebrae; red asterisks indicate sacral vertebrae. A five-allele mutant from the *Hox10* and *Hox11* paralogous mutant group is shown in (**P**) and (**Q**), respectively (red arrows indicate sacral wing formation). Analogous vertebrae were dissected from the control and from each triple mutant to compare single vertebral identities. The 19th vertebral element, normally T12, is shown in (**B**), (**G**), and (**L**). The 23rd element, normally L3, is shown in (**C**), (**H**), and (**M**). The 28th element, normally S2, is

Negative and positive controls in bioinformatics



- **Negative control**: quantify the capability of the program to return a negative answer when there are no regulatory elements.
  - Artificial sequences
    - RSAT ***random-sequences*** (Markov models to mimic k-mer frequencies of the organism)
  - Biological sequences without common regulation
    - RSAT ***random-genes*** (negative control for expression clusters)
    - RSAT ***random-genome-fragments*** (negative controls for ChIP-seq)
  - Randomized motifs: column permutations preserve nucleotide frequencies and information content
    - RSAT ***permute-matrix***
- **Positive control**: quantify the capability of the program to detect known regulatory elements
  - Annotated sites (e.g. sites from TRANSFAC) in their original context (promoter sequences).
  - Annotated sites implanted in other context
    - Biological sequences (random selection).
    - Artificial sequences.
  - Artificial sites implanted in artificial sequences.
    - RSAT ***random-motif***
    - RSAT ***random-sites***
    - RSAT ***implant-sites***



RSAT random-genome-fragments

- Select a set of fragments with random positions in a given genome, and return their coordinates and/or sequences
- Adapted to chip-seq ?
  - Yes: same number of peaks + same size
  - No: composition of the sequences (nucleotides, k-mers) may change depends on genomic regions
  -
- Complexify the control
  - Make sure no peak is covered
  - Take regions close / far from the peaks
  - Maintain same composition
  - Maintain same dataset size
  - …

Why is it important ?

To prevent this ….

NATURE | BRIEF COMMUNICATION ARISING

Universality of core promoter elements?

Matthias Siebert & Johannes Söding

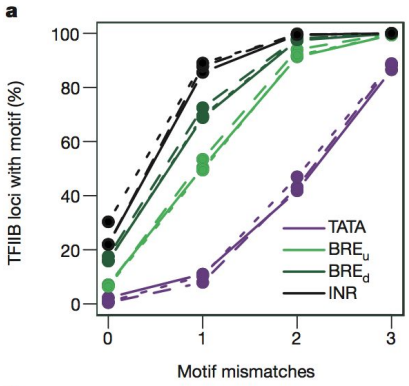
Affiliations | Contributions | Corresponding author

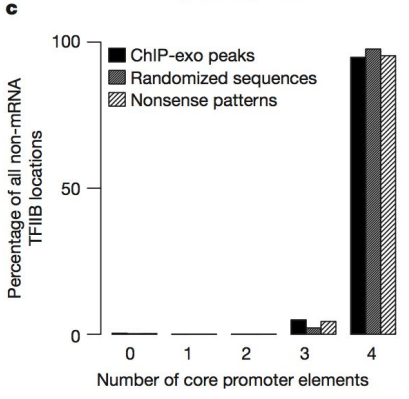
Nature **511**, E11–E12 (24 July 2014) | doi:10.1038/nature13587
Received 06 December 2013 | Accepted 12 June 2014 | Published online 23 July 2014
Retraction (September, 2014)

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

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

We show that the claimed universality of CPEs is explained by the low specificities of the patterns used and that the same match frequencies are obtained with two negative controls (randomized sequences and scrambled patterns).
Our analyses also cast doubt on the biological significance of most of the 150,753 non-messenger-RNA-associated ChIP-exo peaks, 72% of which lie within repetitive regions.





nature International weekly journal of science

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

Archive | Volume 513 | Issue 7518 | Retractions | Article

NATURE | RETRACTION

Retraction: Genomic organization of human transcription initiation complexes

Bryan J. Venters & B. Franklin Pugh

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

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

Subject terms: Transcriptional regulatory elements

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

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

Contact

Matthieu Defrance <matthieu.dc.defrance@ulb.ac.be>

Celine HERNANDEZ <chernand@biologie.ens.fr>,

Stephanie LE GRAS <slegras@igbmc.fr>,

Rachel Legendre <rachel.legendre@pasteur.fr>

Denis Puthier < denis.puthier@uni-amu.fr>,

Morgane THOMAS-CHOLLIER <mthomas@biologie.ens.fr>,

Claire Rioualen <claire.rioualen@inserm.fr>,

Jacques van Helden <Jacques.van-Helden@univ-amu.fr>,

Tao YE <yetao@igbmc.fr>,

Supplementary information

To go further

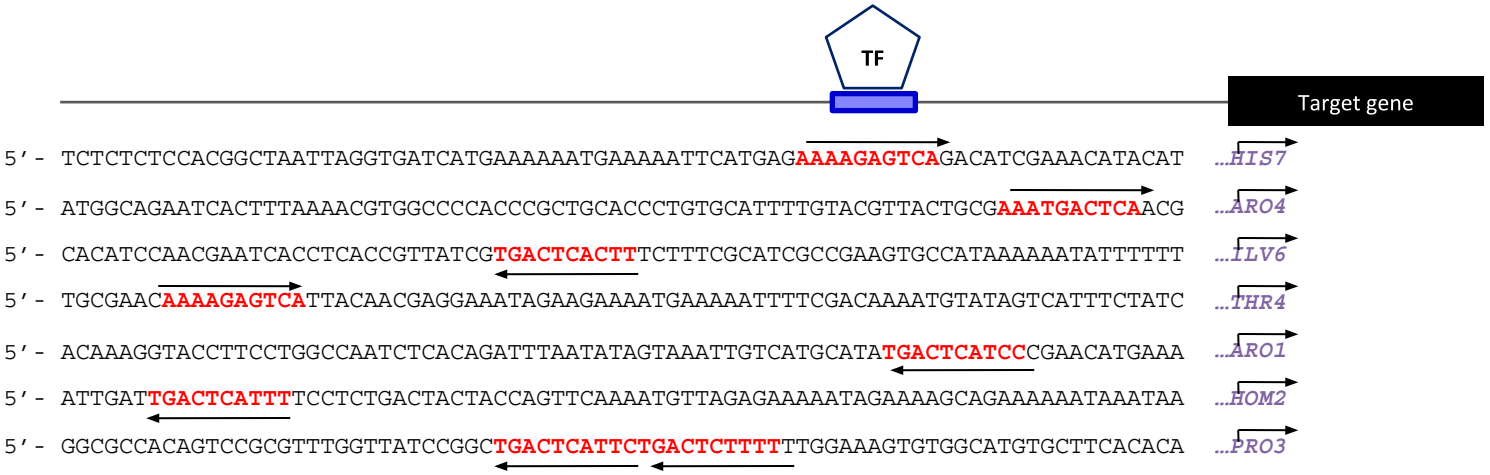
- The next slides explain step by step the algorithm behind oligo-analysis
- Peak-motifs : follow this protocol to grasp the detailed tweaking of parameters (send us an email to have free access to the PDF if necessary)
  - Thomas-Chollier et al. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nature Protocols 7, 1551–1568 (2012).
- Description and evaluation of peak-motifs
  - Matrix-quality : RSAT program that can be used to evaluate the enrichment of motifs in peaks
- Description of the RSAT software suite
  - 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.
- Tutorial for ECCB 2014 : http://rsat.ulb.ac.be/eccb14/

More info: RSAT descriptions + protocols

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

Principle: detect unexpected patterns

- Binding sites are represented as “words” = “oligonucleotides”=“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 (k-mers, oligonucleotides).



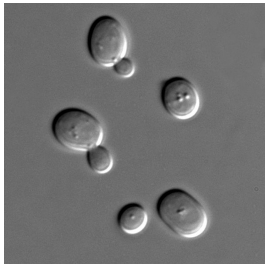
Motif discovery using word counting

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|ttttttc	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|ttttttc	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|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

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|ttttttc	40	tatata|tatata	22
gaaaaa|ttttttc	13	tatata|tatata	40	ataaga|tcttat	20
tgccaa|ttggca	12	aaaaat|attttt	35	aagaaa|tttctt	20
aaaaat|attttt	12	aagaaa|tttctt	29	gaaaaa|ttttttc	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

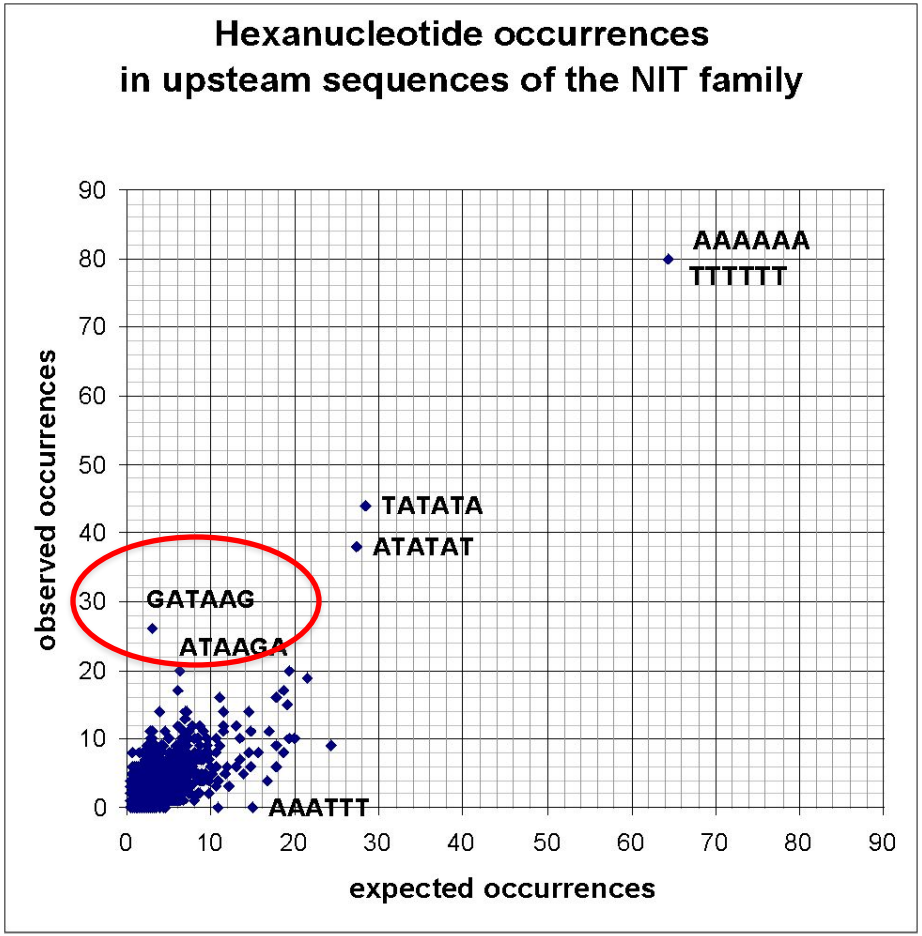
- theoretical background model (Markov Models)

Example:

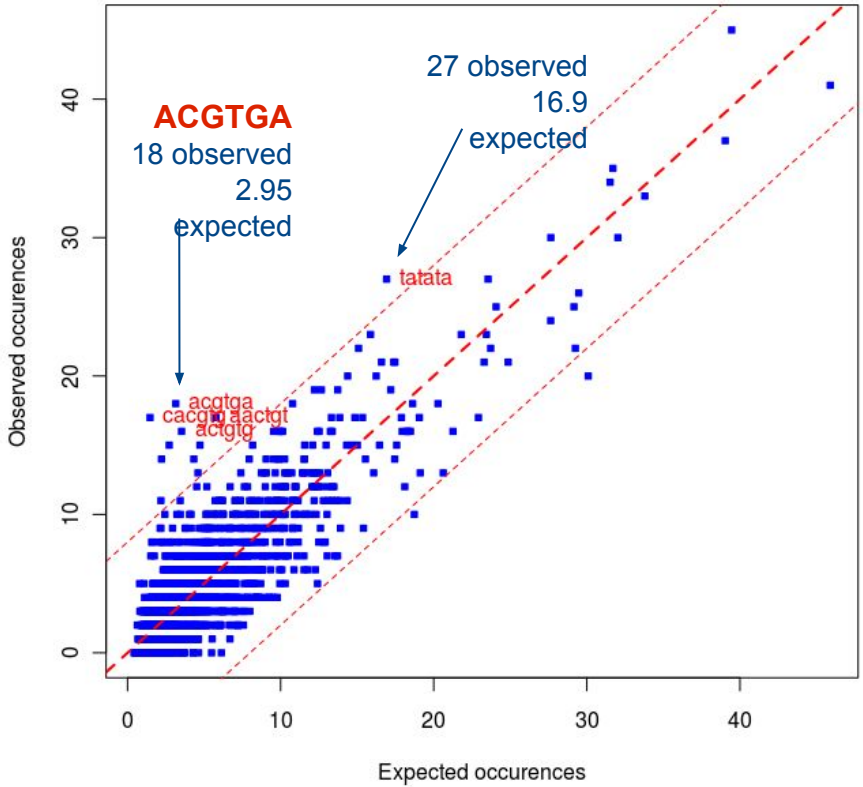
6nt frequencies in the whole set of 6000 yeast **upstream** sequences

;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



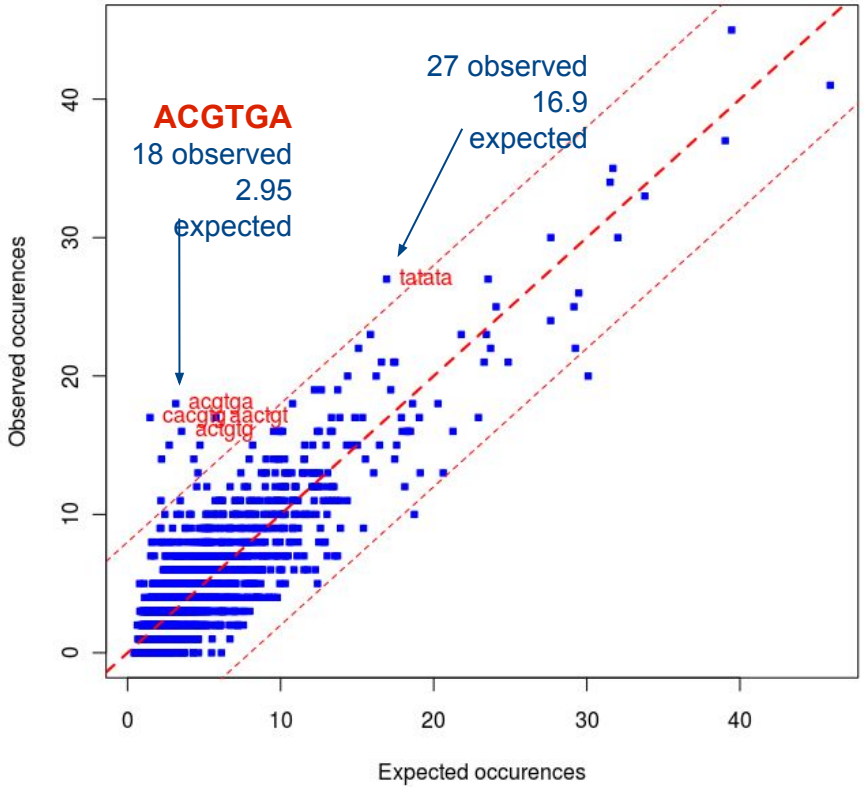
How to evaluate expected number of occurrences ?

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 significance



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

Statistical significance

How « big » is the surprise to observe 18 occurrences when expecting 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 occurrences

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

Binomial distribution to measure the exceptionality of the occurrences

Sequencing

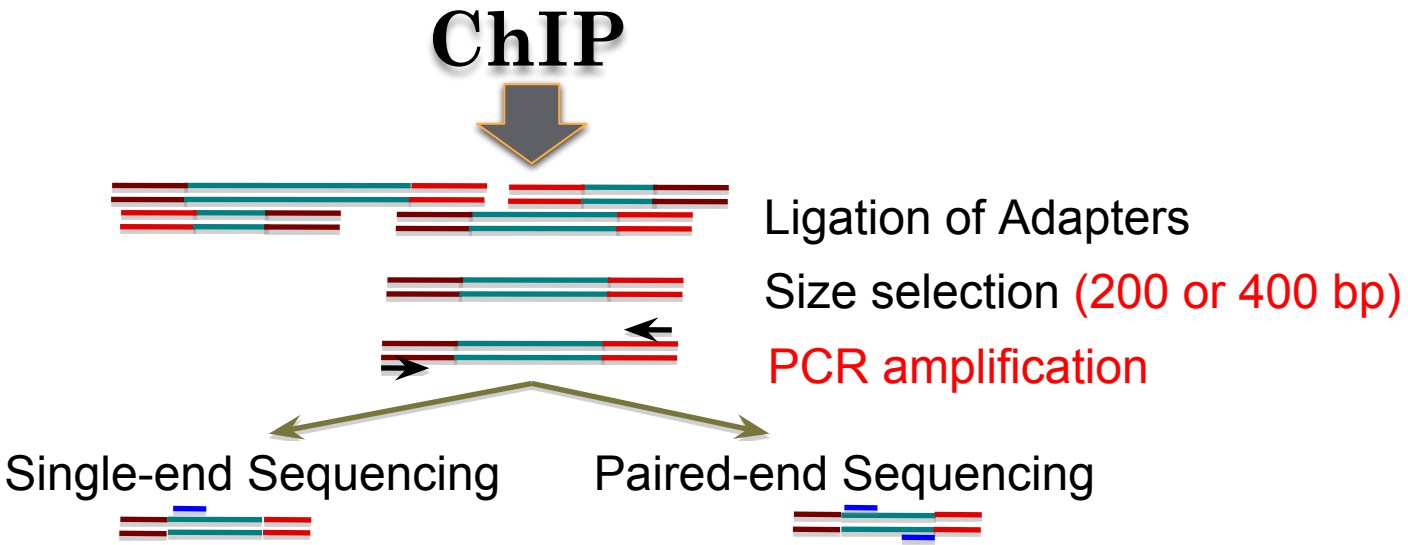
- Sequencer : Illumina HiSeq 4000
- No. of reads per run, per sample :
  - 1st run on the GAIIx : 10-20 millions of reads per lane
  - (HiSeq 2500) 4 samples per lane :~41 millions per sample
  - (HiSeq 4000) 8 samples per lane :~43 millions per sample
- Length of DNA fragment : ~200bp
- No. of cycle per run : 50

Single end or paired end?

- Single end (most of the time)
- Paired-end sequencing
  - Improve identification of duplicated reads
  - Better estimation of the fragment size distribution
  - Increase the mapping efficiency to **repeat regions**
  - The price!

Library prep

- Step between ChIP and sequencing.
- The goal is to prepare DNA for the sequencing.
- Starting material: ChIP sample (1-10ng of sheared DNA).



Considerations on ChIP

- Antibody
  - Antibody quality varies, even between independently prepared batches of the same antibody (Egelhofer, T. A. *et al*. 2011).
- Number of cells
  - Large numbers of cells are required for a ChIP experiment (limitation for small organisms).
- Shearing of DNA (Mnase I, sonication, Covaris): trying to narrow down the size distribution of DNA fragments

Complexity in DNA fragments

Controls

- Used mostly to filter out false positives (high level of noise)
  - Idea: potential false positive will be enriched in both treatment and control.
- A control will fail to filter out false positives if its enrichment profile is very different from the enrichment profile of false positive regions in the treatment sample.
- 3 types of controls are commonly used :
  - ***'Input' DNA***: a portion of DNA sample removed prior to IP
  - ***DNA from non specific IP***: DNA obtained from IP with an antibody not known to be involved in DNA binding or chromatin modification, such as IgG.
  - ***Mock IP DNA***: DNA obtained from IP without antibodies.
- 'Input' most generally prefered.

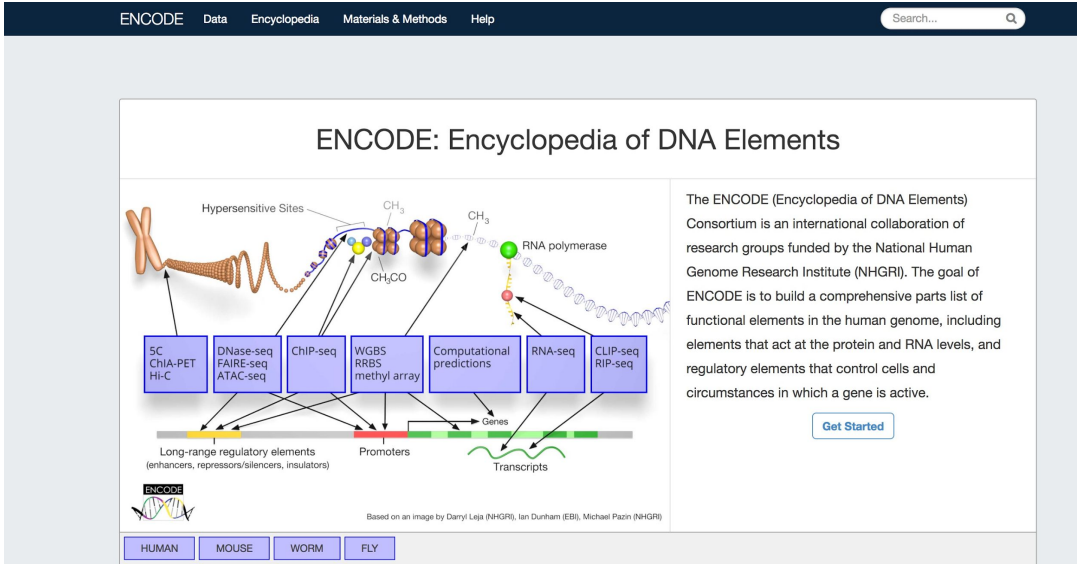
Replicates

- A **minimum** of two replicates should be carried out per experiment.
- Get ***biological replicates*** rather than technical replicates
    - i.e. taken from an independent cell culture, embryo pool or tissue sample.

ENCODE

See: https://www.encodeproject.org/

- The **ENCyclopedia Of DNA Elements** (ENCODE) consortium has carried out hundreds of ChIP-seq experiments and has used this experience to develop a set of working standards and guidelines.



https://www.encodeproject.org/

Protocol

For the sake of time and to avoid repetitions of processing steps already covered in other tutorials, **we have already performed quality-check of the reads and mapping**. **We will thus start from a BAM file**.

Perform the following steps to Import **the BAM files** that will be subsequently analyzed.

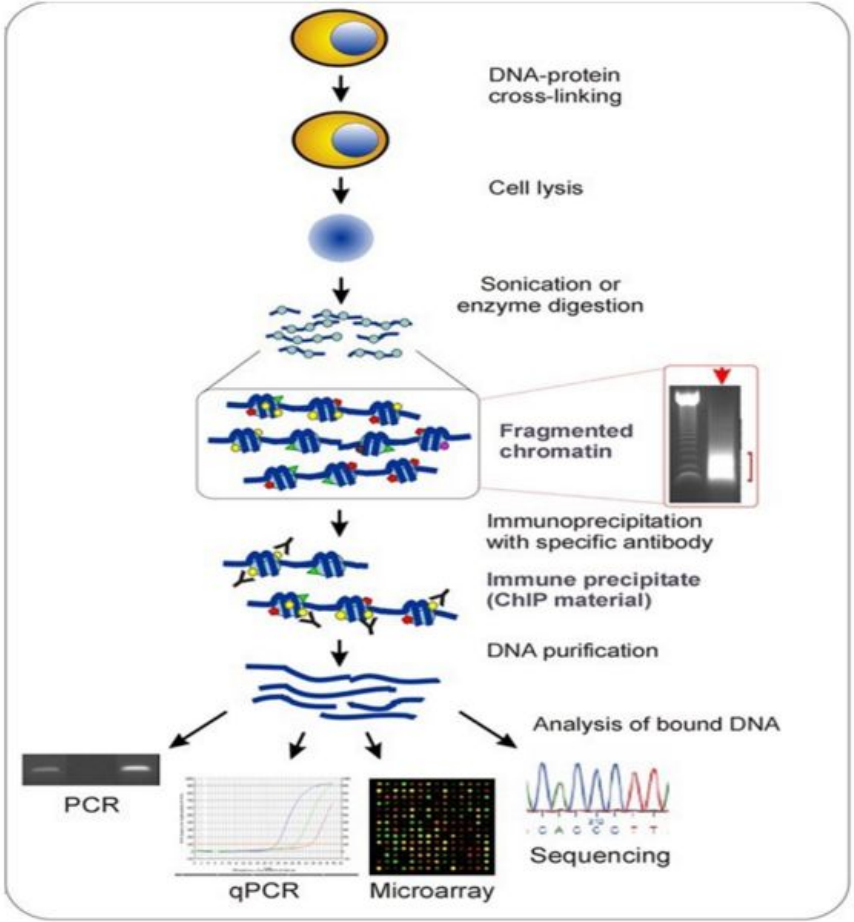
1. Go to **Shared data > data libraries > CHIPSEQ_EBA_2016 > BAM files**
2. Select the files:
   a. ESR1_chr1.bam
   b. input_chr1.bam
3. Click on **to History**
4. Enter the name **ESR1** to create a new history
5. Click on **Import**
6. Click on **Analyze Data** or on **Galaxy** in the top menu to go to the newly created history.

Slide to be removed, see next one

Introduction

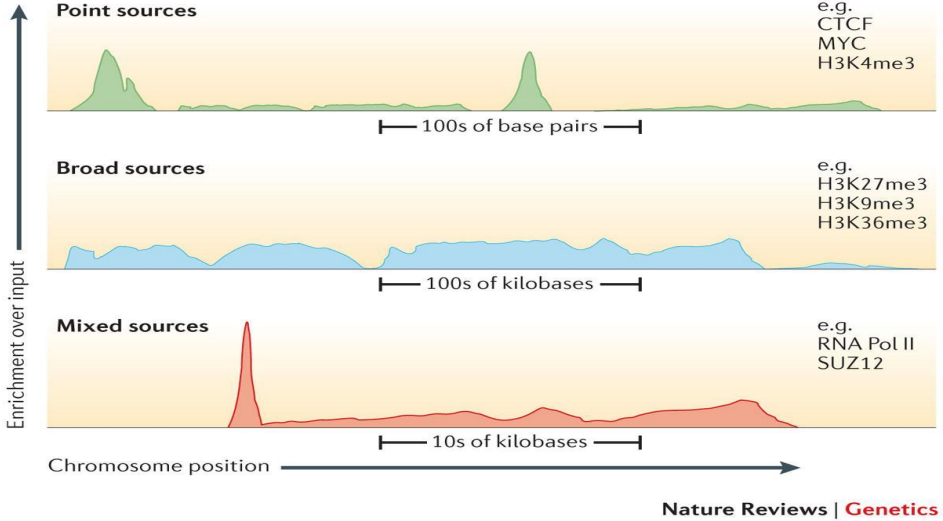
ChIP-seq

- Used to analyze
  - Transcription factor locations.
  - Histone modifications along genome.



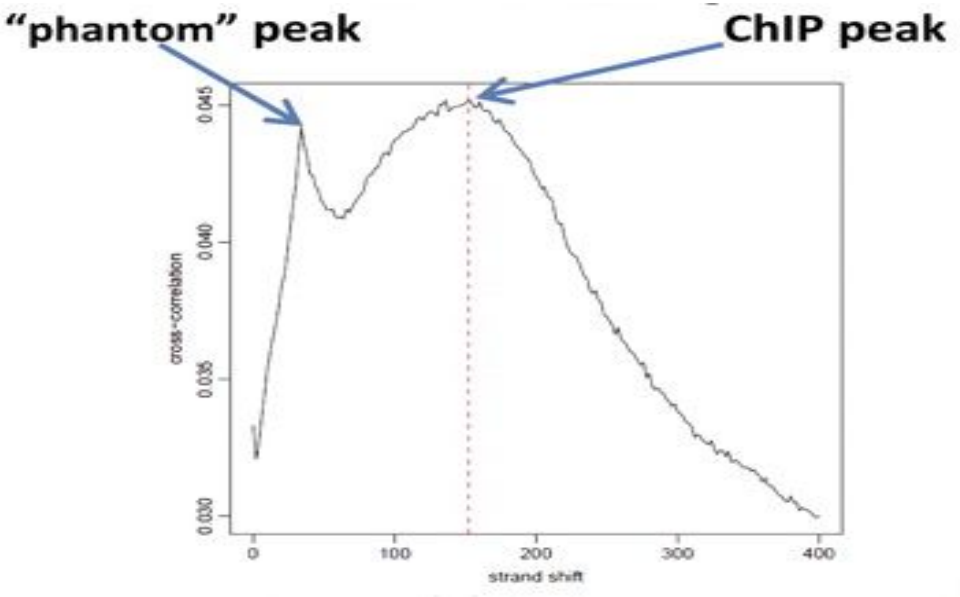
Sequencing depth

- Estimate the required depth depending on:
  - ChIP-ped protein
  - Expected profile type
  - Expected number of binding sites
  - Genome size
- Examples
  - For human genome
    - 20 million uniquely mapped read sequences for point-source peaks.
    - 40 million for broad-source peaks.
  - For fly genome: 8 million reads.
  - For worm genome: 10 million reads.

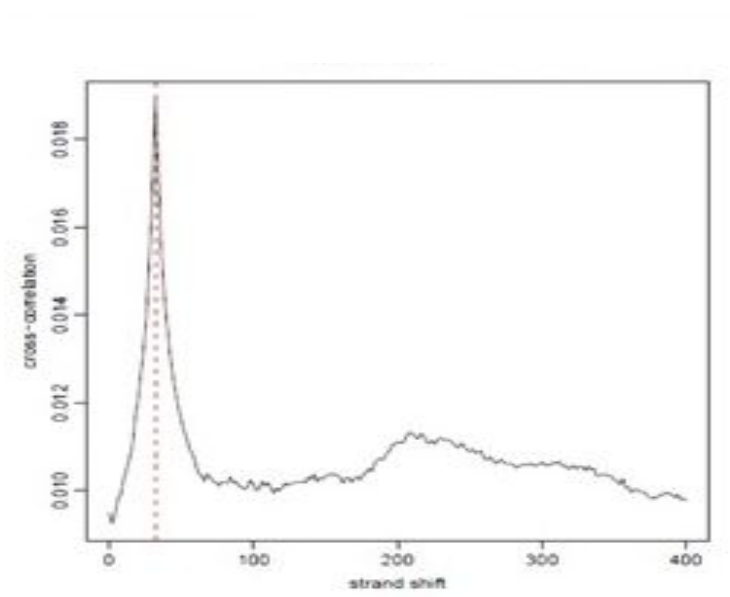


QC: Strand cross-correlation

Successful



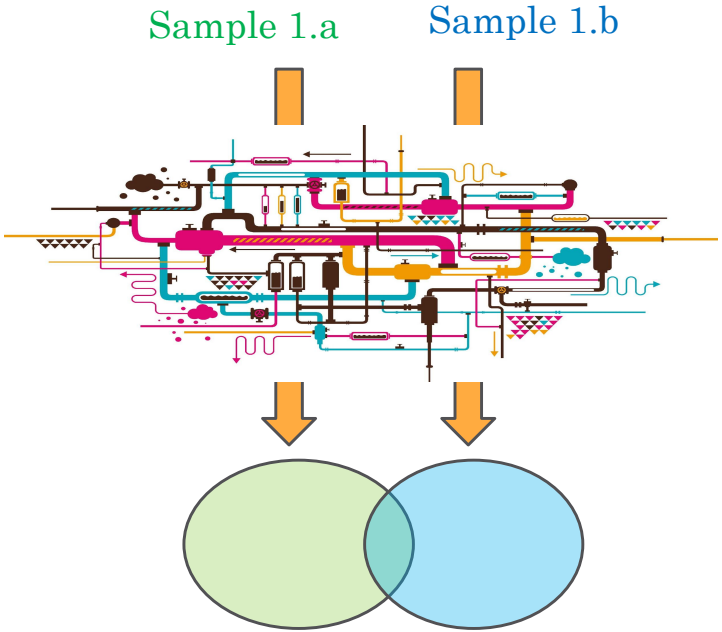
Failed



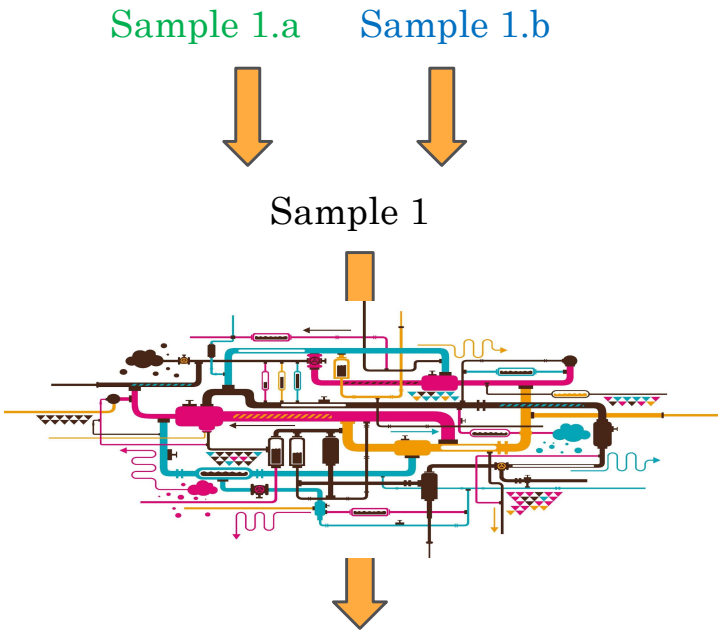
How to deal with replicates?

How to deal with replicates

Analyze samples separately and take union or intersection of resulting peaks



Merge samples prior to the peak calling (e.g recommended by MACS)

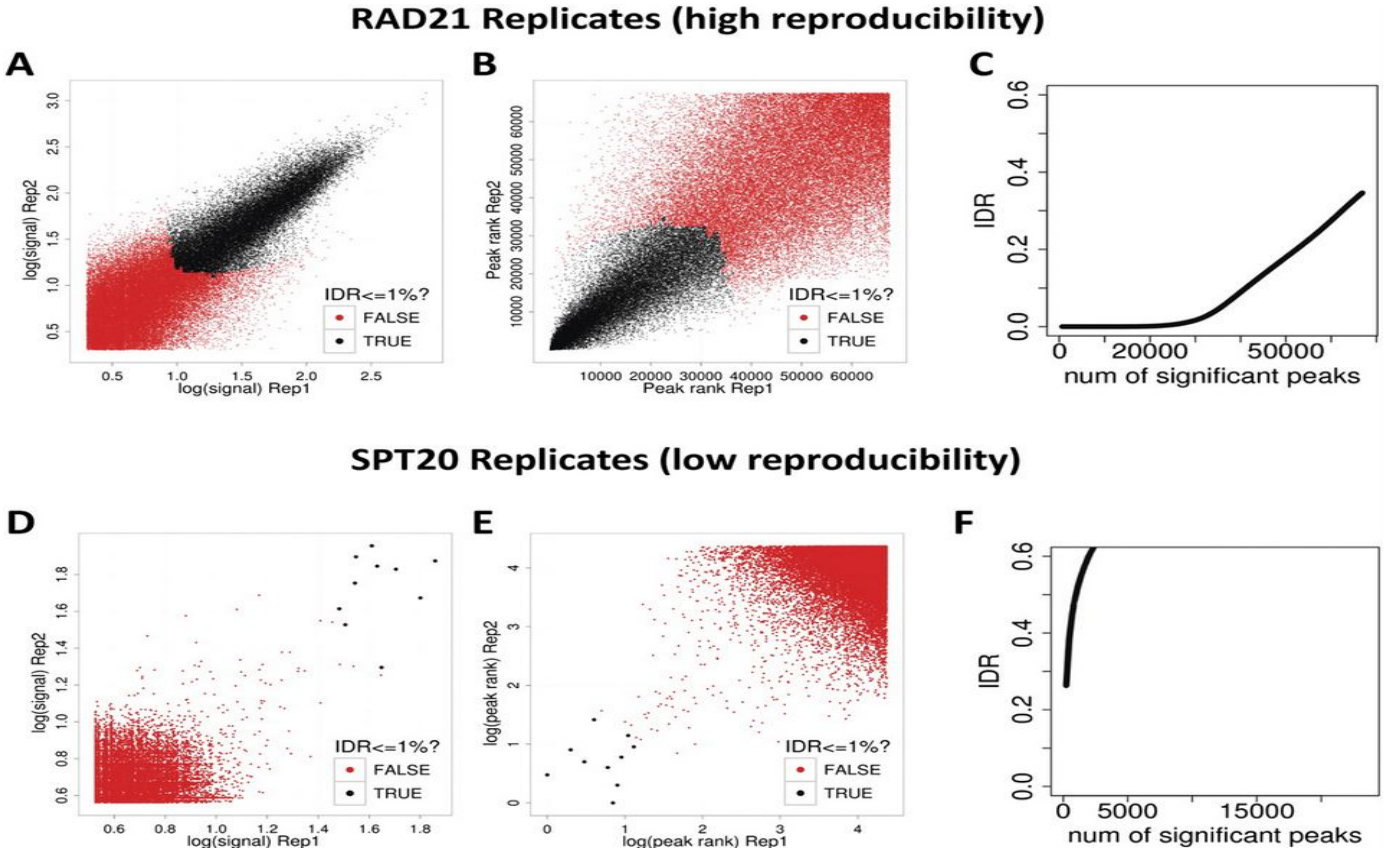


IDR

- IDR = Irreproducible Discovery Rate.
- Measures (in)consistency between replicates.
- Uses reproducibility between score rankings of peaks in the respective replicates to determine an optimal cutoff for significance.
- Idea:
  - The most significant peaks are expected to have high consistency between replicates.
  - The peaks with low significance are expected to have low consistency.

https://sites.google.com/site/anshulkundaje/projects/idr

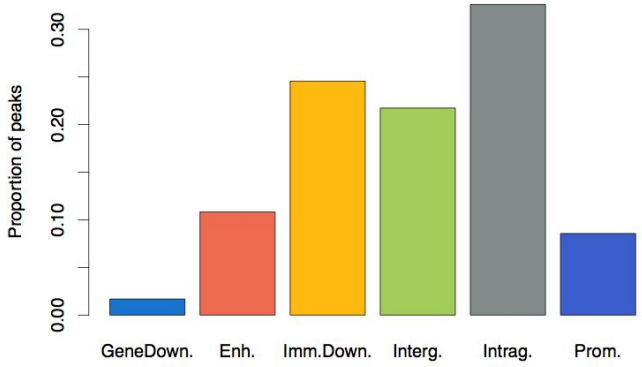
IDR



(!) IDR doesn't work on broad source data!

Galaxy: Annotate peaks

- Input
  - bed file with peaks
- Output
  - Fraction of peaks per genomic elements and annotated peaks



Chromosome	Start	End	Max	Score	DistTSS	Type	TypeIntra
chr1	3001827	3002328	3002077	55.28	659502	intergenic	NA
chr1	3067471	3067948	3067709	50.67	593870	intergenic	NA
chr1	3660316	3662844	3661580	352.43	-1	promoter	NA
chr1	3842462	3842994	3842728	59.21	-181149	intergenic	NA
chr1	3877254	3877710	3877482	52.72	-215903	intergenic	NA
chr1	3939314	3939679	3939496	82.99	-277917	intergenic	NA
chr1	4206037	4206512	4206274	50.86	144121	intergenic	NA
chr1	4481463	4484213	4482838	268.57	3656	intragenic	intron
chr1	4486799	4487684	4487241	88.18	-747	promoter	NA
chr1	4561258	4562489	4561873	236.23	-75379	intergenic	NA
chr1	4635092	4635552	4635322	52.32	140485	intergenic	NA
chr1	4760253	4761284	4760768	111.13	15039	5kbDownstream	NA
chr1	4773759	4776746	4775252	540.12	555	immediateDownstream	f_intron
chr1	4797157	4800182	4798669	249.77	696	immediateDownstream	intron
chr1	4841219	4842788	4842003	156.84	-6405	enhancer	NA
chr1	4846807	4849844	4848325	377.92	-83	promoter	NA
chr1	4873314	4873950	4873632	66.94	25224	intragenic	intron
chr1	4885079	4885564	4885321	64.12	36913	intragenic	intron

Protocol

Relate peaks to genes

In this step we will compute the number of peaks intersecting genomic feature. This approach is very close to the job performed by CEAS. But wait, we will see...

Procedure

1. Search **AnnotatePeaks** in the Galaxy toolbox.
2. Select the **ESR1_peaks.bed** file for analysis.
3. Select **hg19** as genome version and leave other parameters with default values.

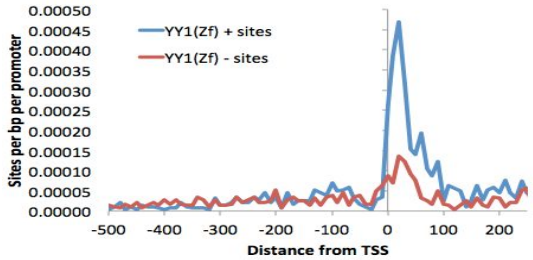
Q1: Compare to the results obtained from CEAS analysis would you draw the same conclusion about promoter regions? What is the importance of running a statistical test in this context?

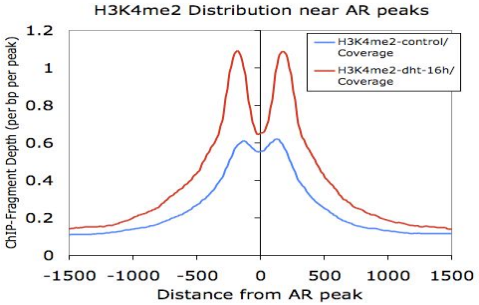
HOMER

Motif discovery and NGS data analysis

Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities

Sven Heinz,[1,7] Christopher Benner,[1,7] Nathanael Spann,[1,7] Eric Bertolino,[4] Yin C. Lin,[3] Peter Laslo,[6] Jason X. Cheng,[4] Cornelis Murre,[3] Harinder Singh,[4,5] and Christopher K. Glass[1,2,*]





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185	79670	Hs.597057	NM_001185	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1|	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195	145282	Hs.660396	NM_001195	ENSG000001	MIPOL1	DKFZp313M	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembra
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

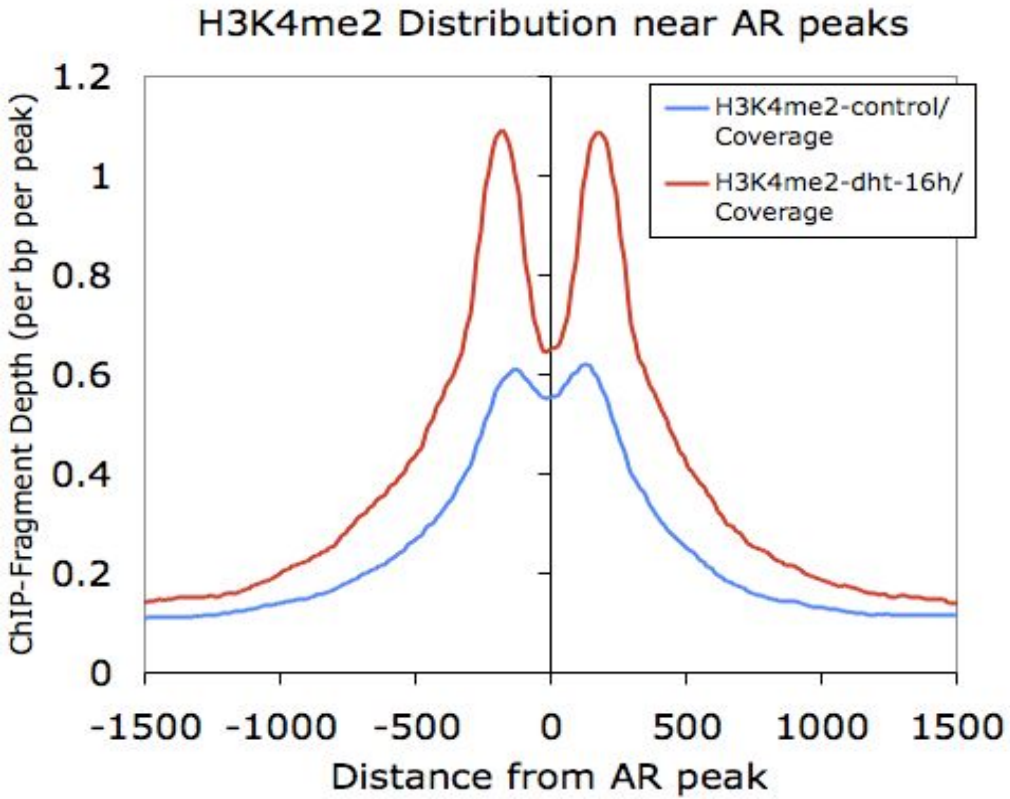
http://homer.salk.edu/homer/

HOMER: annotate peaks

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185	79670	Hs.597057	NM_001185	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195	145282	Hs.660396	NM_001195	ENSG000001	MIPOL1	DKFZp313M2	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembr
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

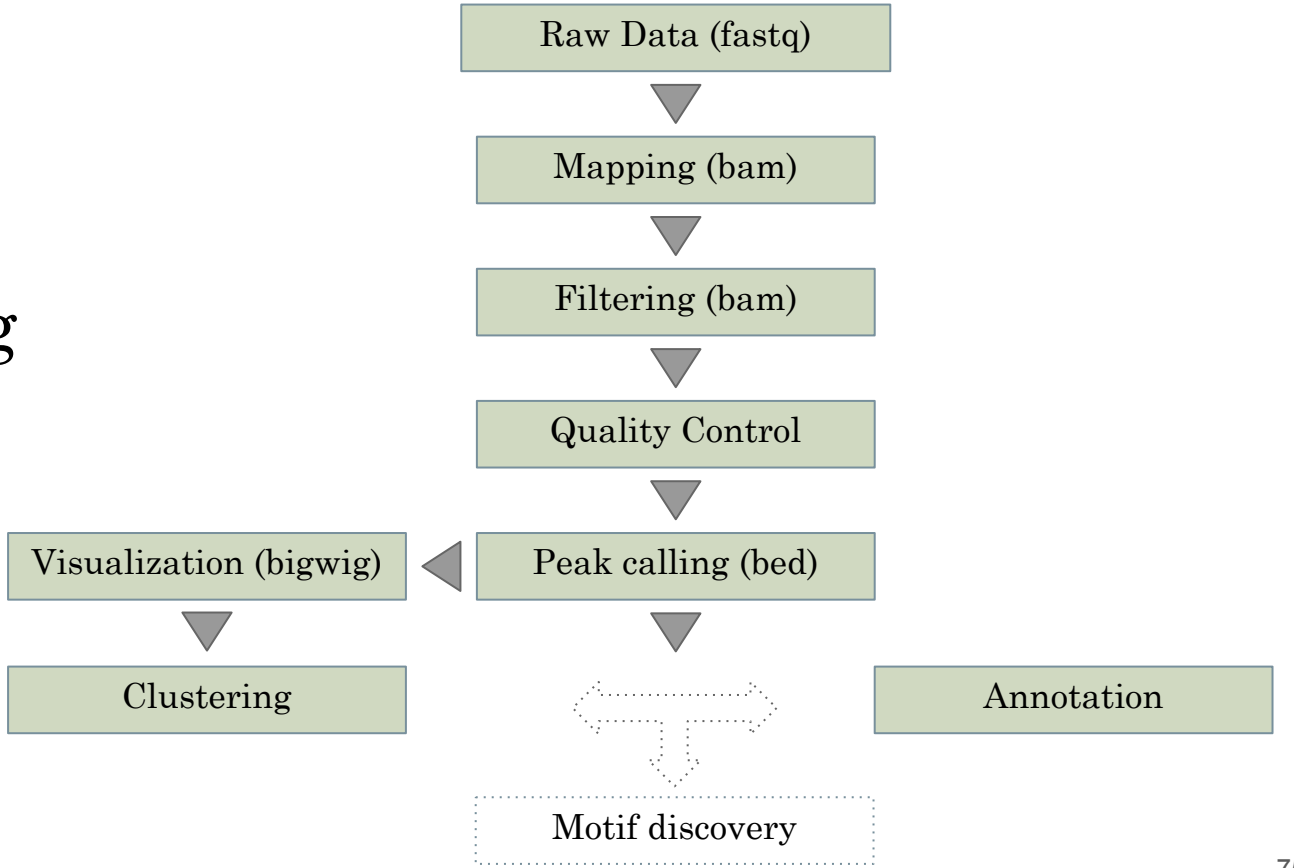
1 Peak ID
2 Chromosome
3 Peak start position
4 Peak end position
5 Strand
6 Peak Score
7 FDR/Peak Focus Ratio/Region Size
8 Annotation (i.e. Exon, Intron, ...)
9 Detailed Annotation (Exon, Intron etc. + CpG Islands, repeats, etc.)
10 Distance to nearest RefSeq TSS
11 Nearest TSS: Native ID of annotation file
12 Nearest TSS: Entrez Gene ID
13 Nearest TSS: Unigene ID
14 Nearest TSS: RefSeq ID
15 Nearest TSS: Ensembl ID
16 Nearest TSS: Gene Symbol
17 Nearest TSS: Gene Aliases
18 Nearest TSS: Gene description
19 Additional columns depend on options selected when running the program.

HOMER: compare peaks



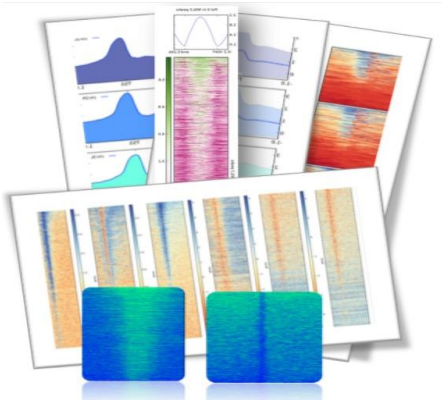
Peak co-occurrence statistics
Co-bound peaks
Differentially bound peaks

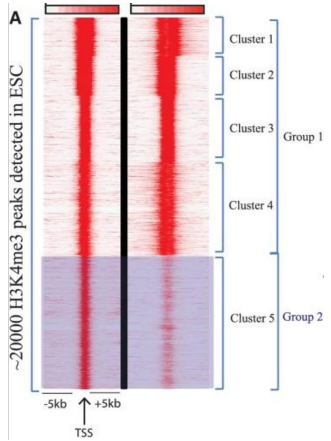
Clustering


Raw Data (fastq)
Mapping (bam)
Filtering (bam)
Quality Control
Peak calling (bed)
Visualization (bigwig)
Clustering
Annotation
Motif discovery

Based on signal distribution, are there any classes of genomic regions?

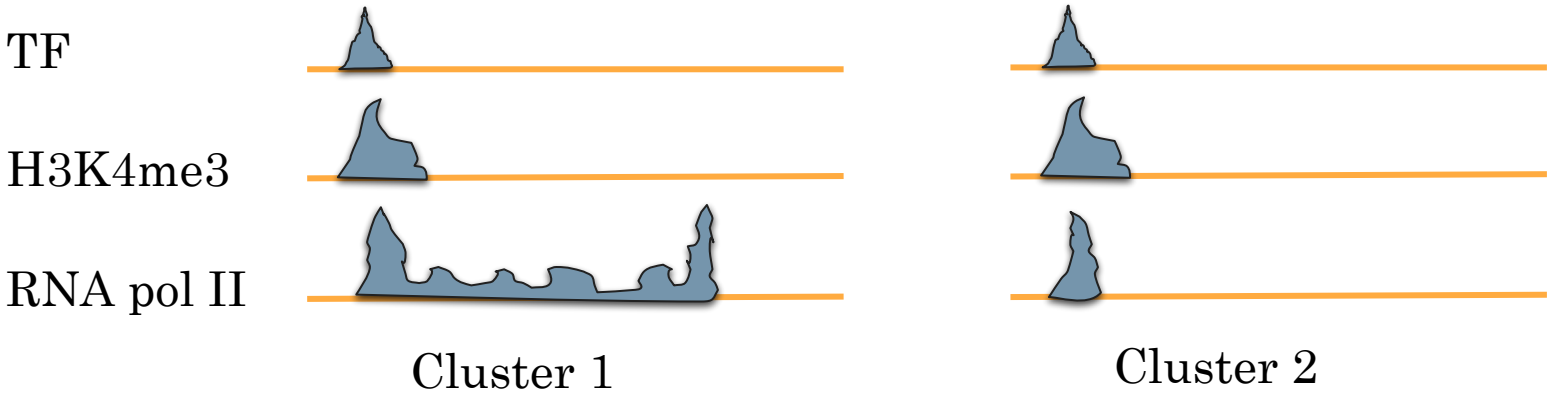
- How does the signal (read counts) distribute around or inside:
  - Transcriptional start sites (TSS)
  - Transcriptional termination sites (TTS)
  - Gene bodies, exons, introns
- Tools:
  - Deeptools (heatmapper)
  - seqMINER
- Unsupervised clustering methods (e.g k-means)
  - Discover some underlying classes of genomic regions





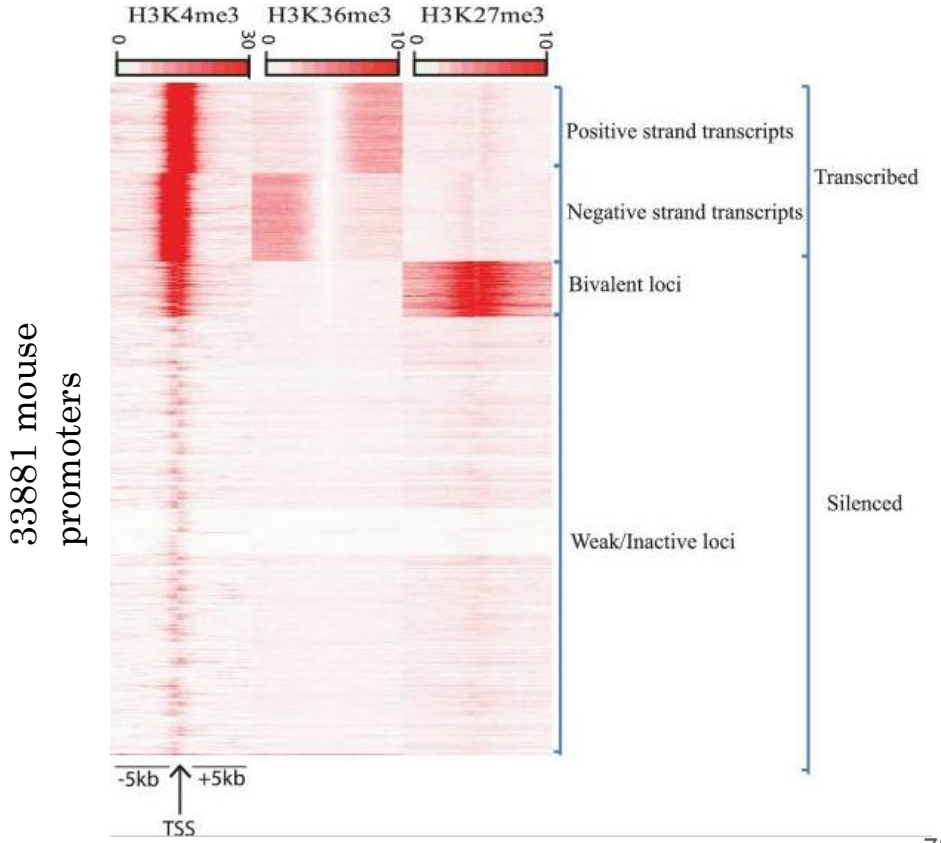
Clustering

- Group together genomic regions with similar enrichments
- In a single sample or multiple samples
- E.g:



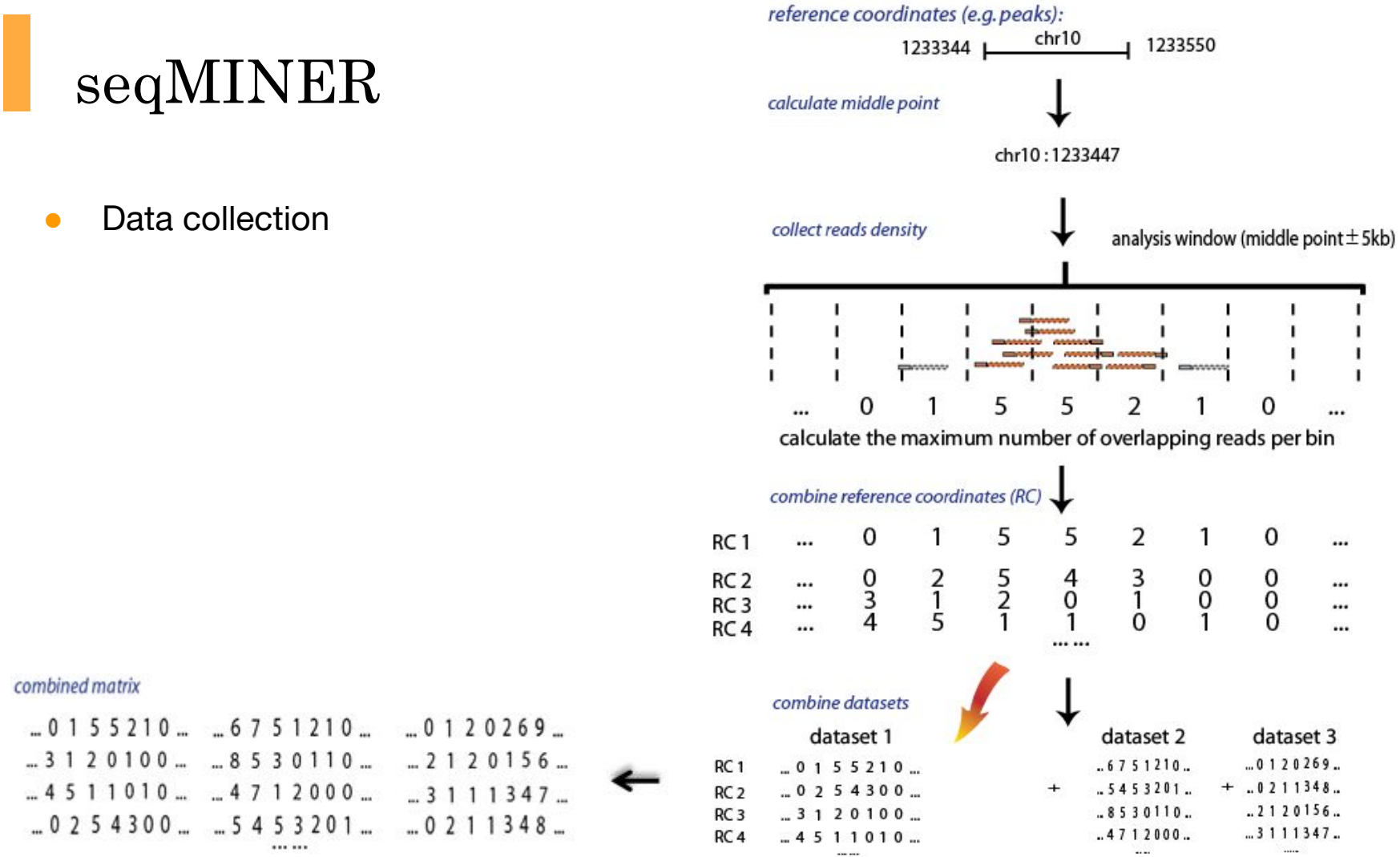
Clustering

- **seqMINER**
  - User friendly interactive interface with multiple graphical representations
  - Multiple dataset comparison
  - Java, multi-platform



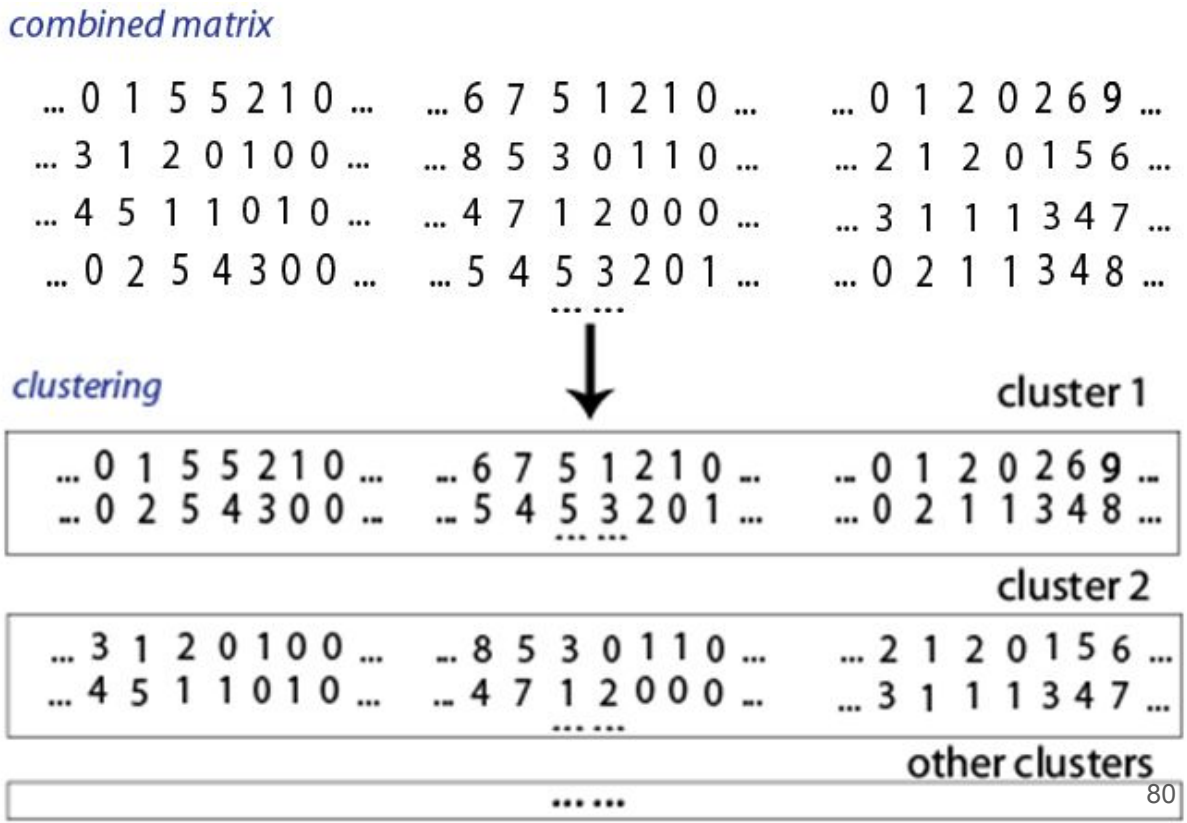
seqMINER

- Data collection



seqMINER

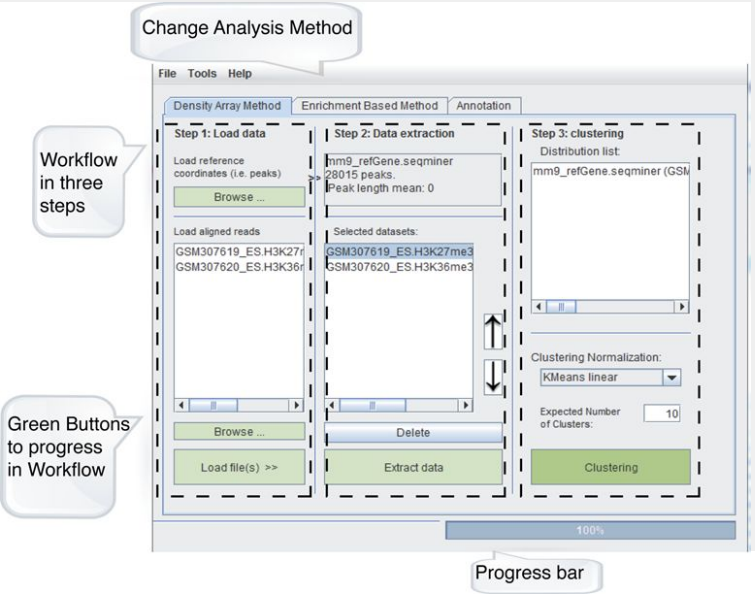
- Clustering (K-means)

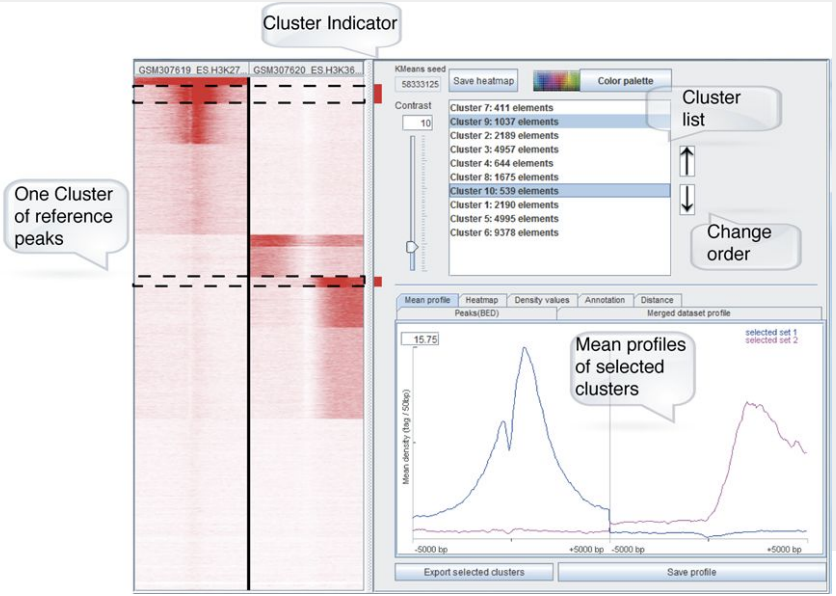


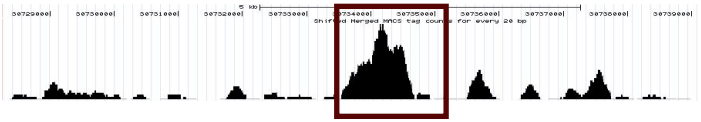
Protocol

seqMINER demonstration

- Download seqMINER from sourceforge
  - https://sourceforge.net/projects/seqminer/files/latest/download



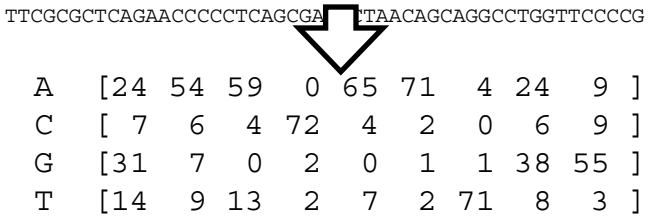




ChIP-seq peaks

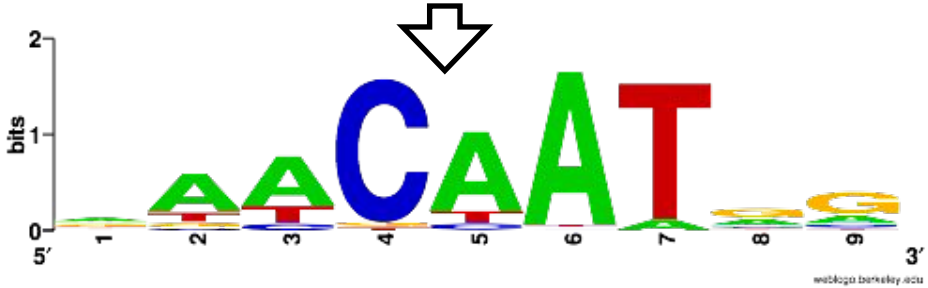
```
>mm9_chr1_39249116_39251316_+
gagaggaaggggagaaagagggaggggagGGTGATAGGTAGCCAGGAG
CCAATGGGGGCGTTTTCCTTGTCCAGGCCACTTGCTGGAATGTGAGATGT
AGAATGACCCAAAGAGAGCTGCCAAGACAGAGCTCTGCCCCAGGAATTGA
ACTCAAAGGGTGTCAGAAAGCAGGTGGCCTTTGTGCACCTGGCGCGGGGA
CGTGGCTCCCCTCTTCCGGCTGGTCTAGCCAGGtgcctgcctgcctgcct
gccGTGATCTCTGGACGCCAGTAGAGGGTTGTTGTGGGTTTGGGTGAAAC
ACGCCACCCCTGAGCTCTTCCGCGGGGCTAGCAATCTCCCCATCACCCCA
TTCGCGCTCAGAACCCCCTCAGCGA TTAACAGCAGGCCTGGTTCCCCG
```

DNA sequence



```
A  [24 54 59  0 65 71  4 24  9 ]
C  [ 7  6  4 72  4  2  0  6  9 ]
G  [31  7  0  2  0  1  1 38 55 ]
T  [14  9 13  2  7  2 71  8  3 ]
```

Discovered motif



Motif logo