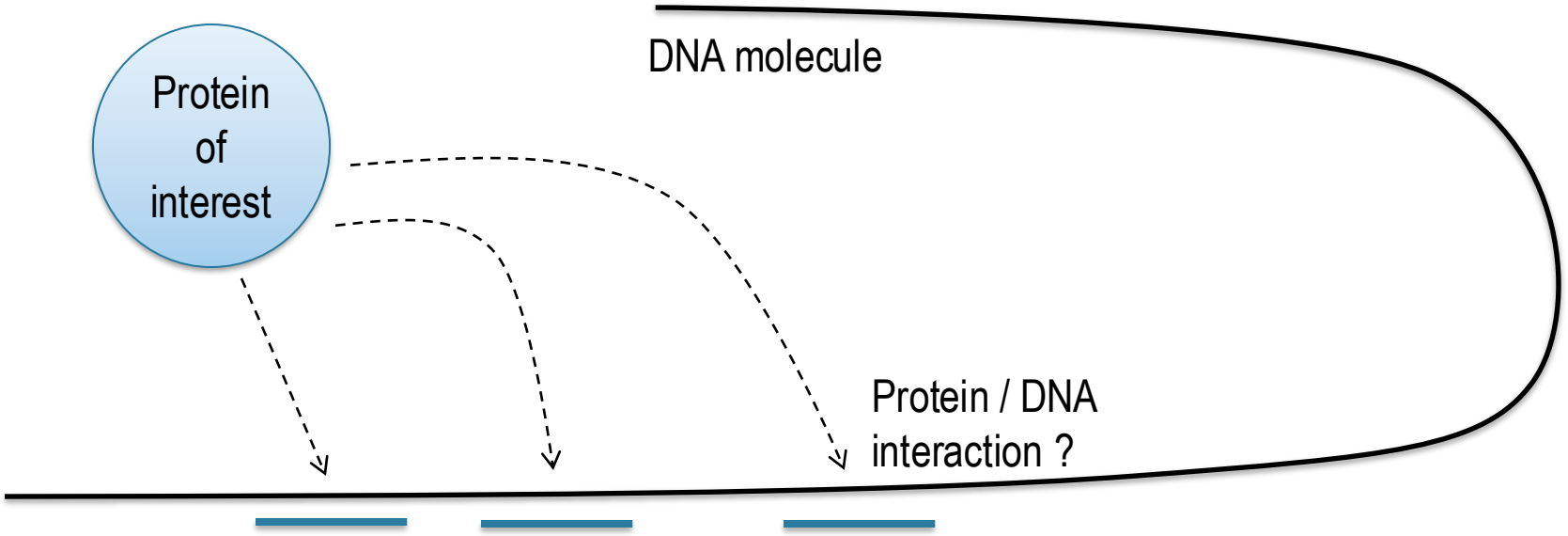
ChIPseq data analysis

GAËLLE LELANDAIS

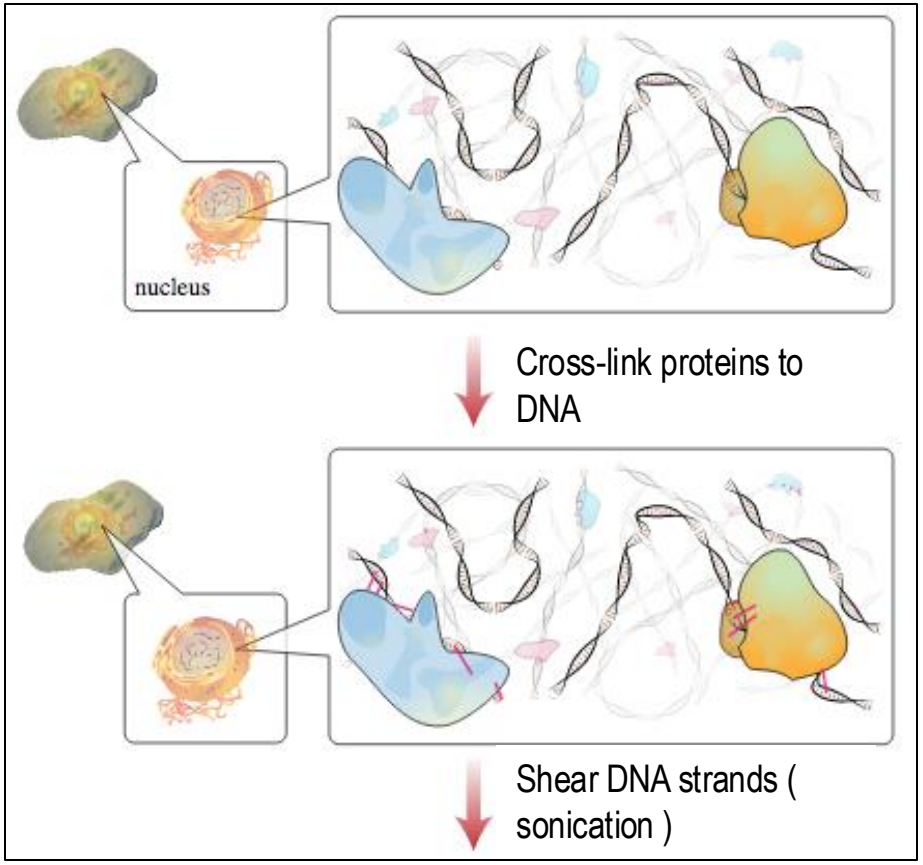
gaelle.lelandais@universite-paris-saclay.fr

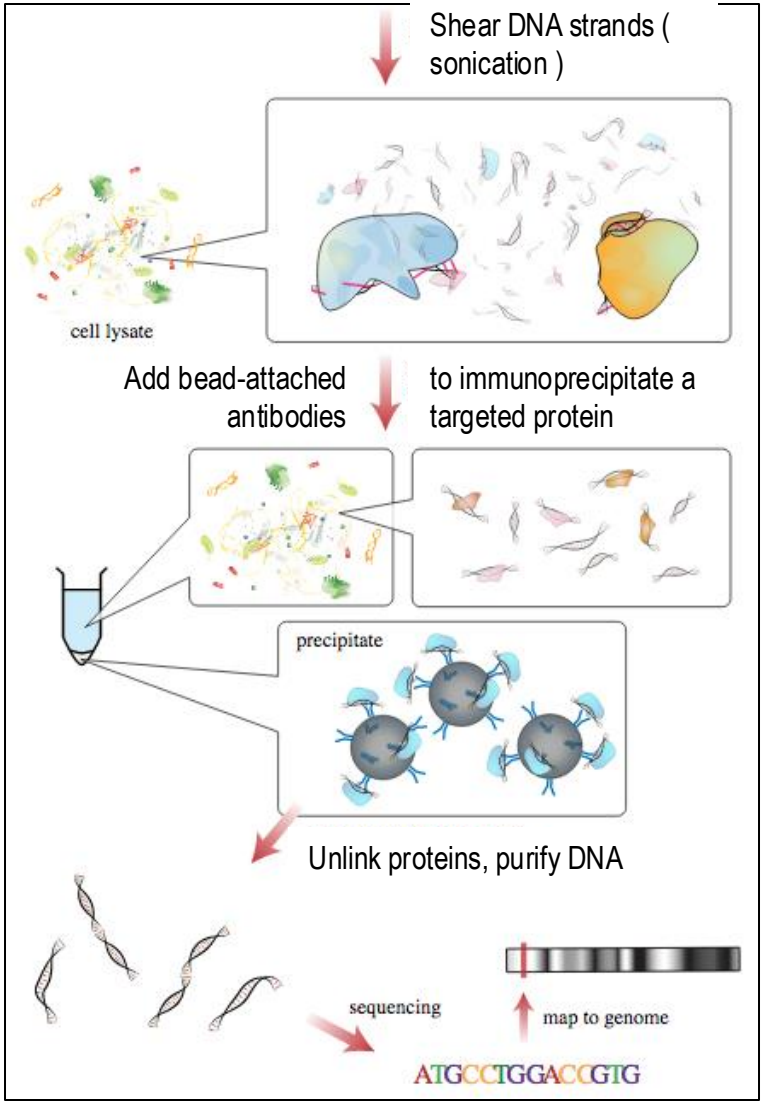
Ultra simplified view of a ChIPseq experiment ☺

Aim : To localize all DNA binding sites for a protein of interest



Experimental protocol



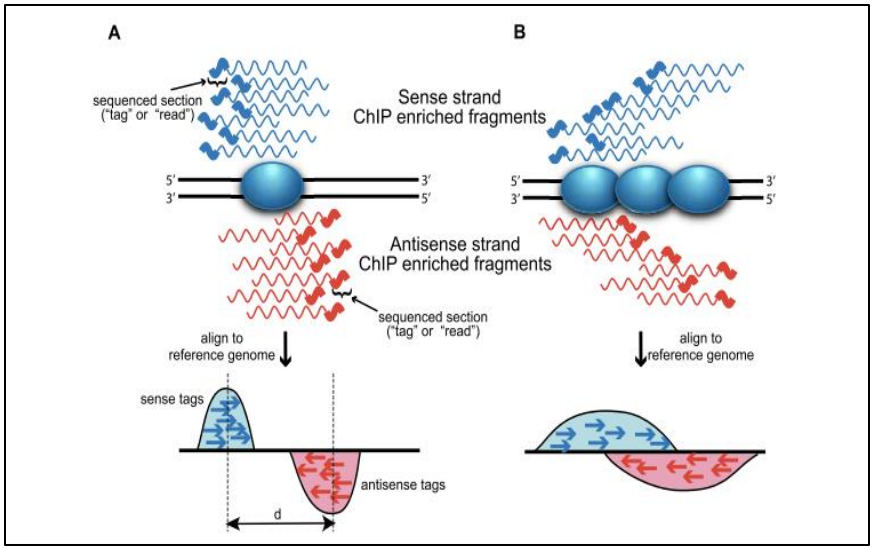


By Jkwchui - Cell diagram adapted from LadyOfHats' Animal Cell diagram. Information based on Illumina data sheet, as well as ChIP and immunoprecipitation articles & references., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17890854

What ChiPseq results look like ?

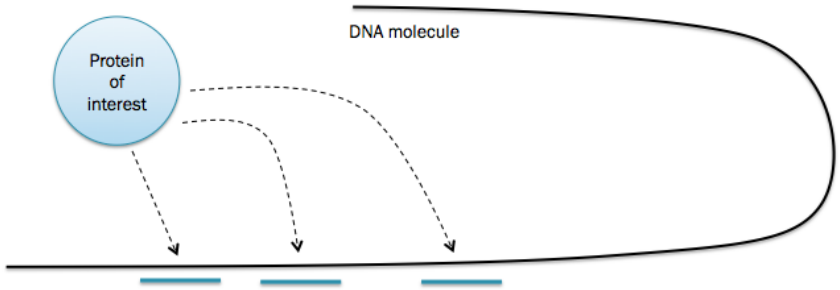
- **In theory, we expect :**

Wilbanks et al. (2010)



these are, what we call “peaks”,

i.e. DNA regions that interact with the protein of interest

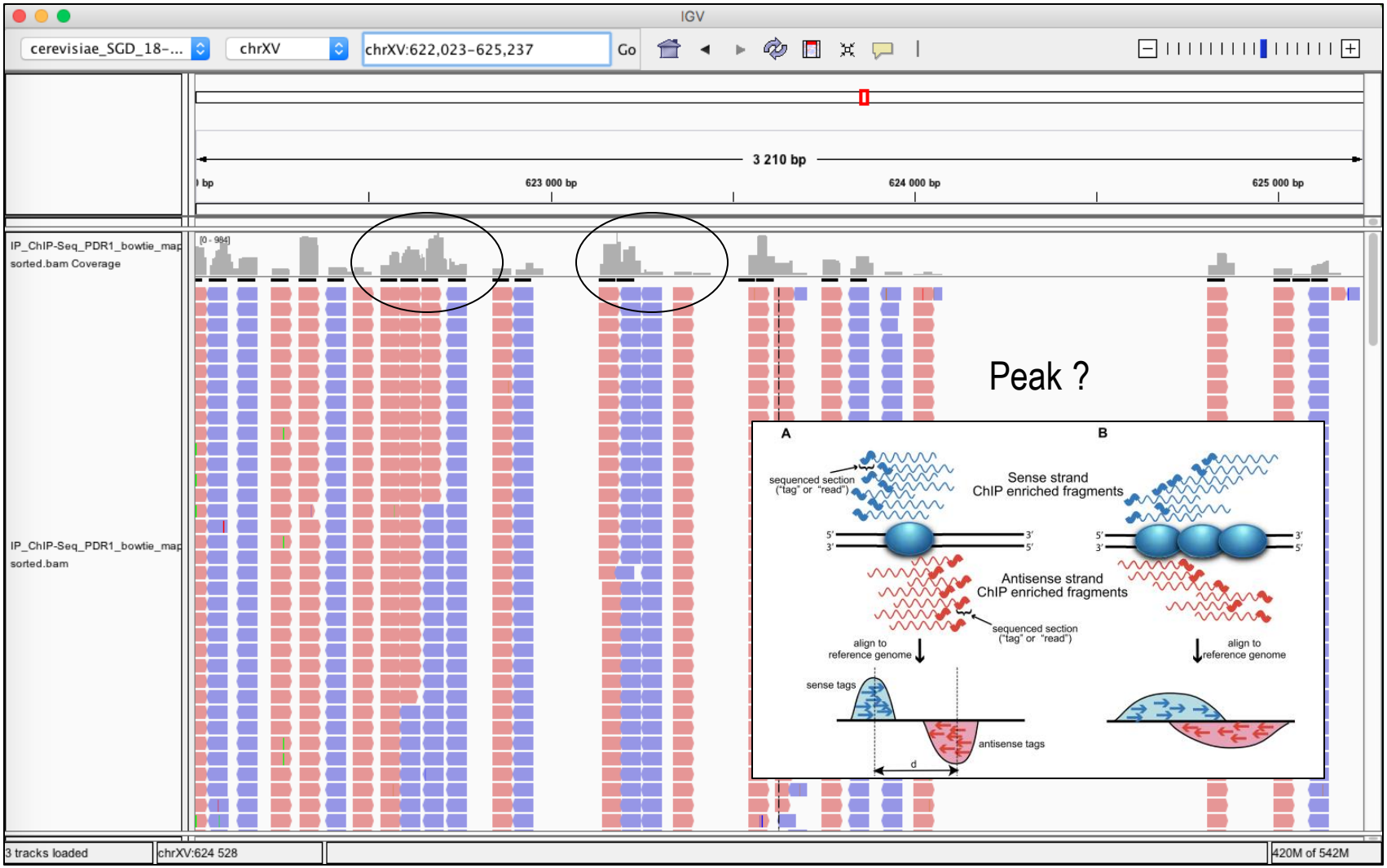


- **Example of ChiPseq results in yeast *S. cerevisiae* (Pdr1p TF) :**

Integrative Genomics Viewer

Peak calling : Search for genomic regions with a high density of reads

IGV screenshot – IP sample



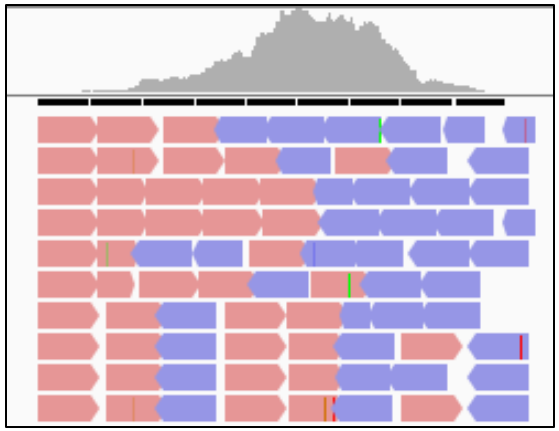
Pdr1p transcription factor in *Saccharomyces cerevisiae*

- Pdr1p belongs to the GAL_4 family of yeast TFs,
- It plays a central role in the regulation of pleiotropic drug resistance through transcriptional controls of about 30 genes,
- Pdr1p is a promoter-resident regulator, which does not need a particular environmental stimulation to bind DNA,
- Several groups have studied the genome-wide binding patterns of Pdr1p using ChIP on chip technology (DeRisi et al., 2000; Devaux et al., 2001; Fardeau et al., 2007),
- The set of genes regulated by Pdr1p has been extensively described in the literature.

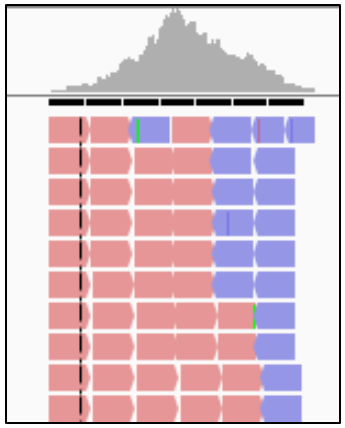
Peaks detected in promoters of Pdr1p target genes

Integrative Genomics Viewer

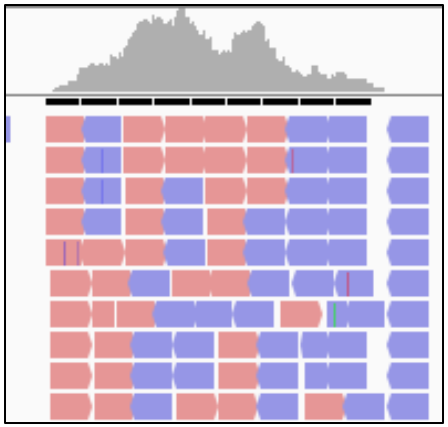
The genes SNQ2, TPO1, PDR5, RPN4 are emblematic Pdr1 targets



SNQ2 promoter



TPO1 promoter

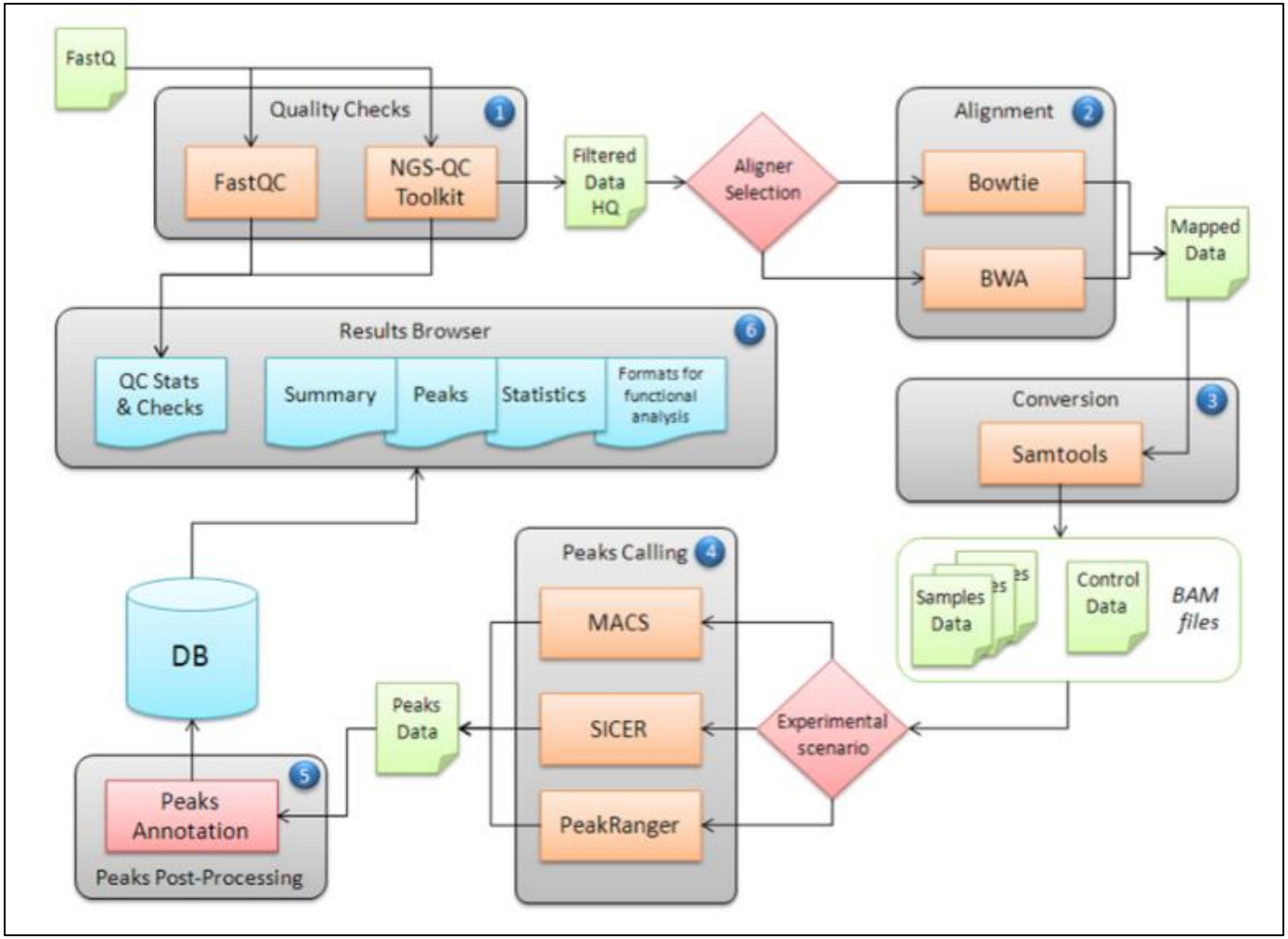


PDR5 promoter

RPN4 promoter

As expected, we can observe “peaks” in each promoter

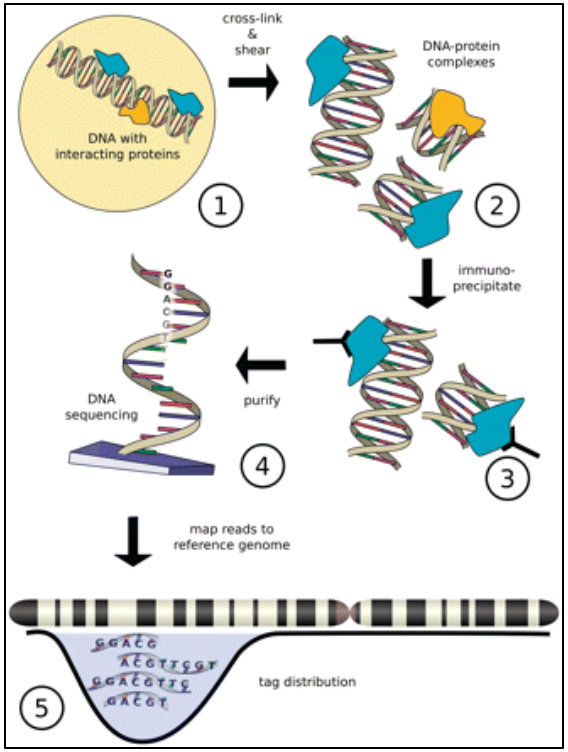
ChIPseq : raw data processing



Olivier KIRSH, M2 Biologie-Informatique (Univ. Paris 7)

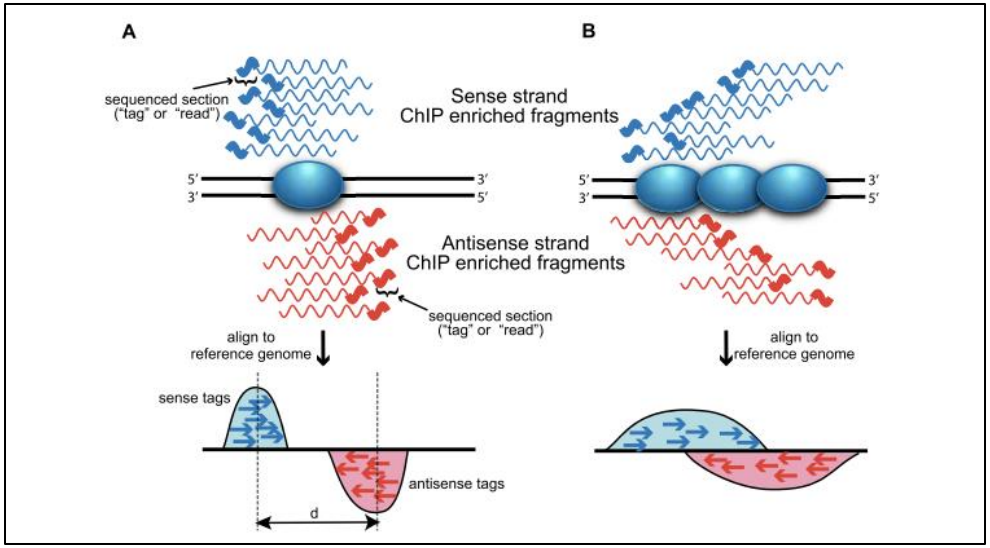
Peak calling, why and how ?

ChIPseq : DNA binding sites of proteins (TFs here)



Szalkowski et al. (2010)

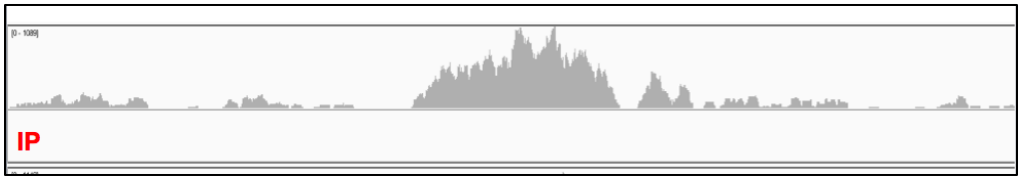
Peak calling : genomic regions with a high density of reads



Wilbanks et al. (2010)

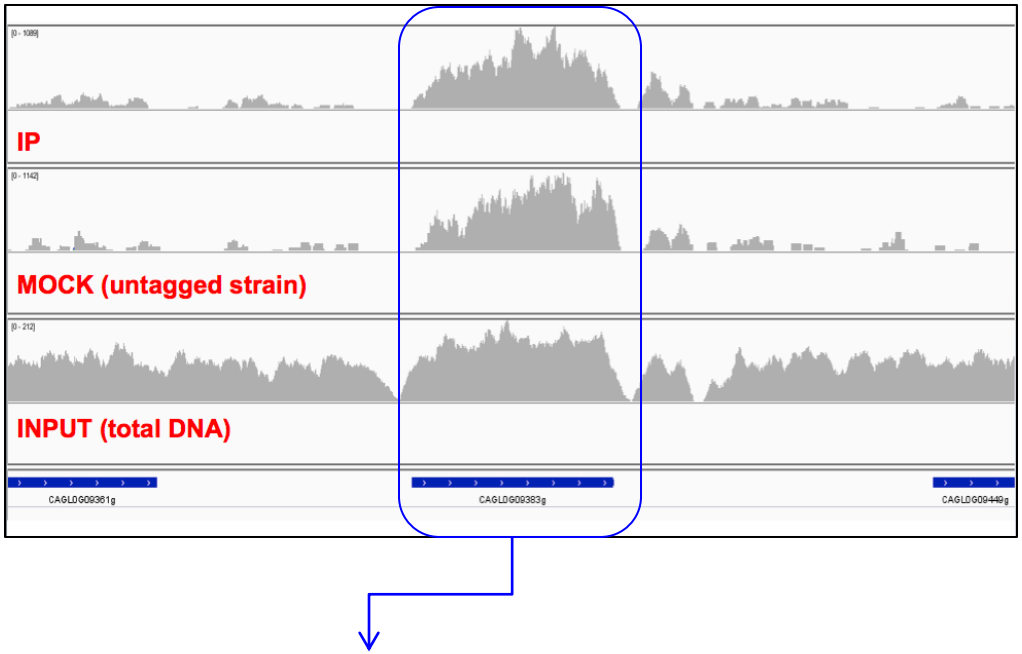
Many computational programs exist to perform peak calling (MACS, SPP, FindPeaks …)

Be careful with the ChIPseq artefacts



Be careful with the ChIPseq artefacts

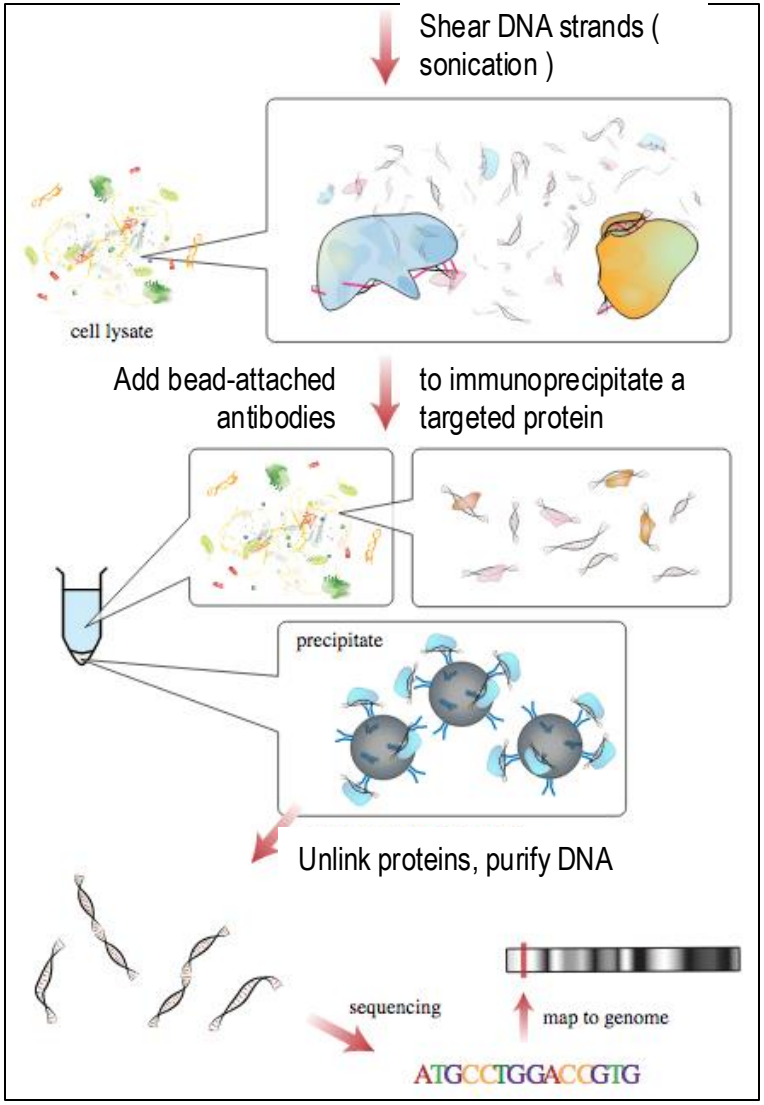
Peaks located in genes, with high expression



Artefact well described in the literature

Any peak calling program will find this peak if the appropriate control is not chosen

Different ChIPseq (classical) controls



An appropriate control data set is critical for analysis of any ChIP-seq experiment because DNA breakage during sonication is not uniform.

1) INPUT control

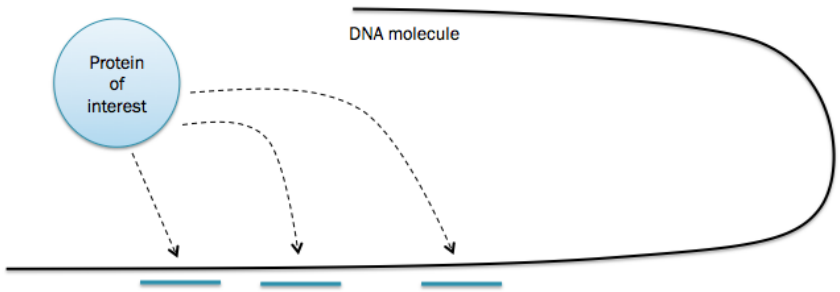
DNA is isolated from cells that have been cross-linked and fragmented under the same conditions as the immunoprecipitated DNA.

2) MOCK control

A ChIP reaction is performed using a control antibody that reacts with an irrelevant antigen.

Peak calling outputs, BED files

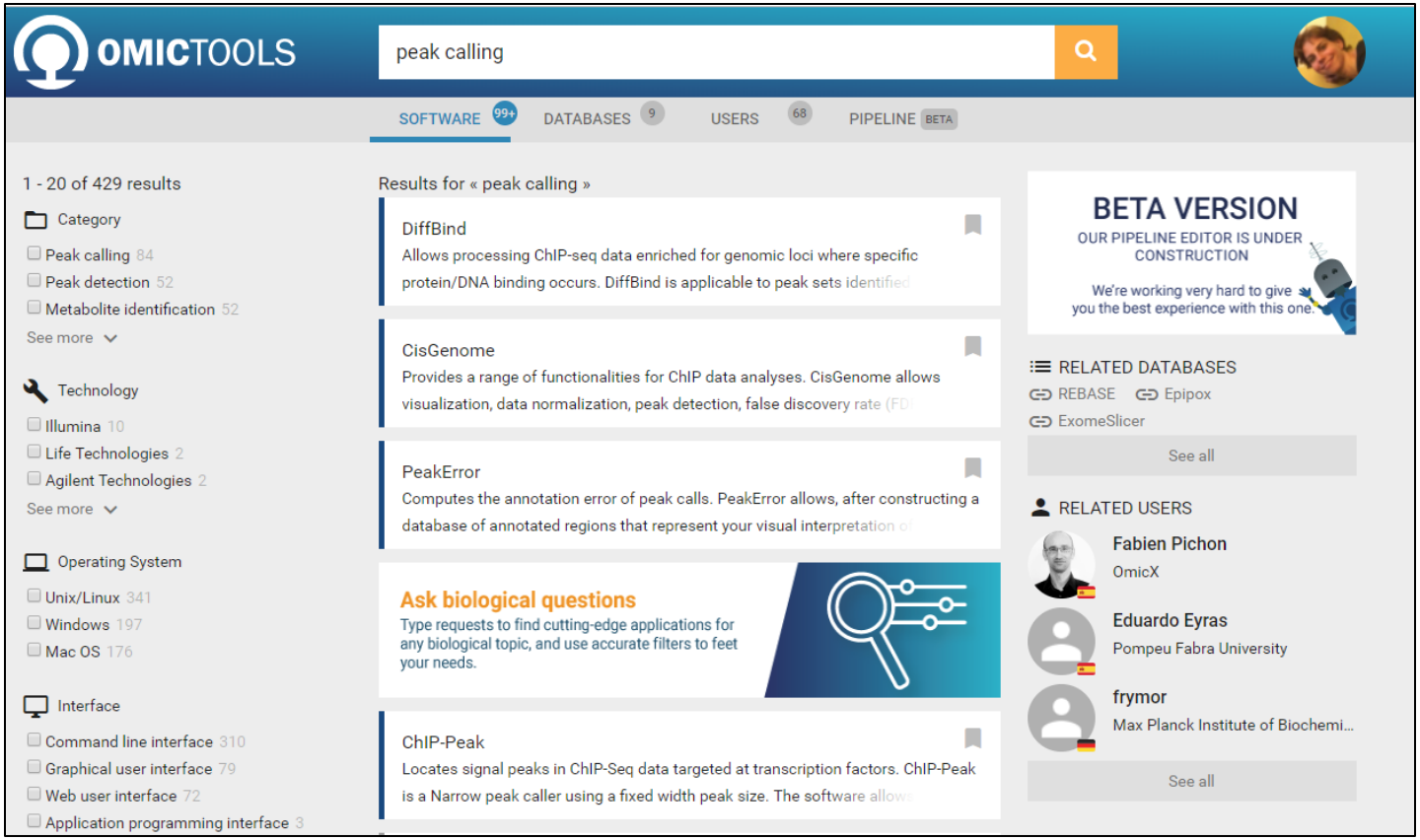
"BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in **an annotation track**. BED lines have **three required fields** and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track." (UCSC : http://genome.ucsc.edu/FAQ/FAQformat#format1)



	Start	End
chr7	127471196	127472363
chr7	127472363	127473530
chr7	127473530	127474697

Peak calling methods, in the literature

➥ Many different analytical programs exist (+99 !)



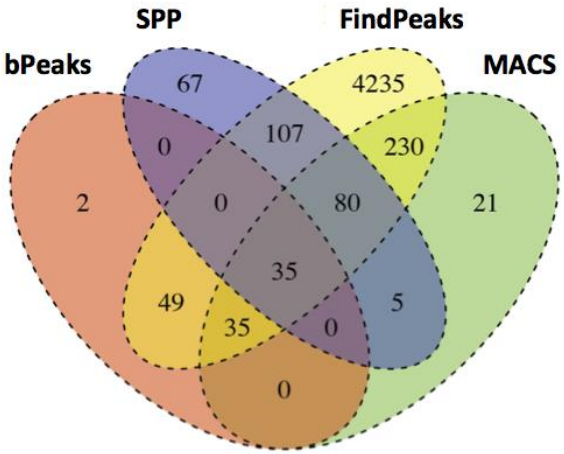
https://omictools.com/

➥ Choosing the correct algorithm and parameter optimized values is a difficult task

Peak calling results, using different algorithms …

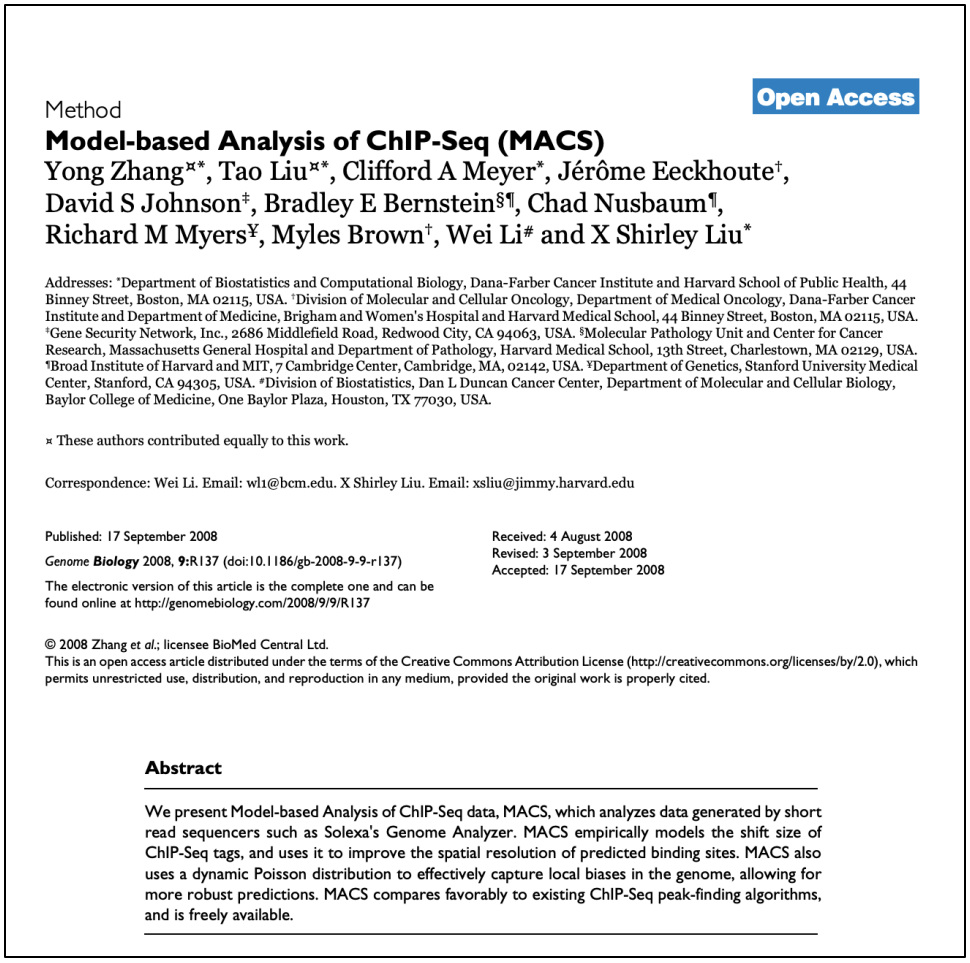
Different computational programs,
used with default parameters :

Method	# of detected peaks	Average peak size (bp)
bPeaks	122	184
SPP	67	2.878
FindPeaks	6.087	739
MACS	248	932



What does it mean ?

The most popular method is...



Method

Open Access

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang¤*, Tao Liu¤*, Clifford A Meyer*, Jérôme Eeckhoute†, David S Johnson‡, Bradley E Bernstein§¶, Chad Nusbaum¶, Richard M Myers¥, Myles Brown†, Wei Li# and X Shirley Liu*

Addresses: *Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. †Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. ‡Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. §Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. ¶Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. ¥Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. #Division of Biostatistics, Dan L Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

¤ These authors contributed equally to this work.

Correspondence: Wei Li. Email: wl1@bcm.edu. X Shirley Liu. Email: xsliu@jimmy.harvard.edu

Published: 17 September 2008

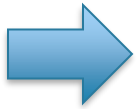
Genome ***Biology*** 2008, **9:**R137 (doi:10.1186/gb-2008-9-9-r137)

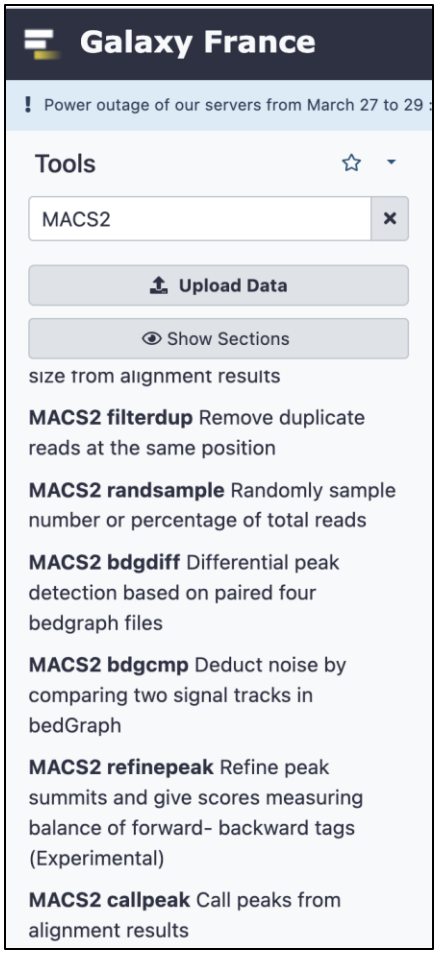
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/R137

Received: 4 August 2008
Revised: 3 September 2008
Accepted: 17 September 2008

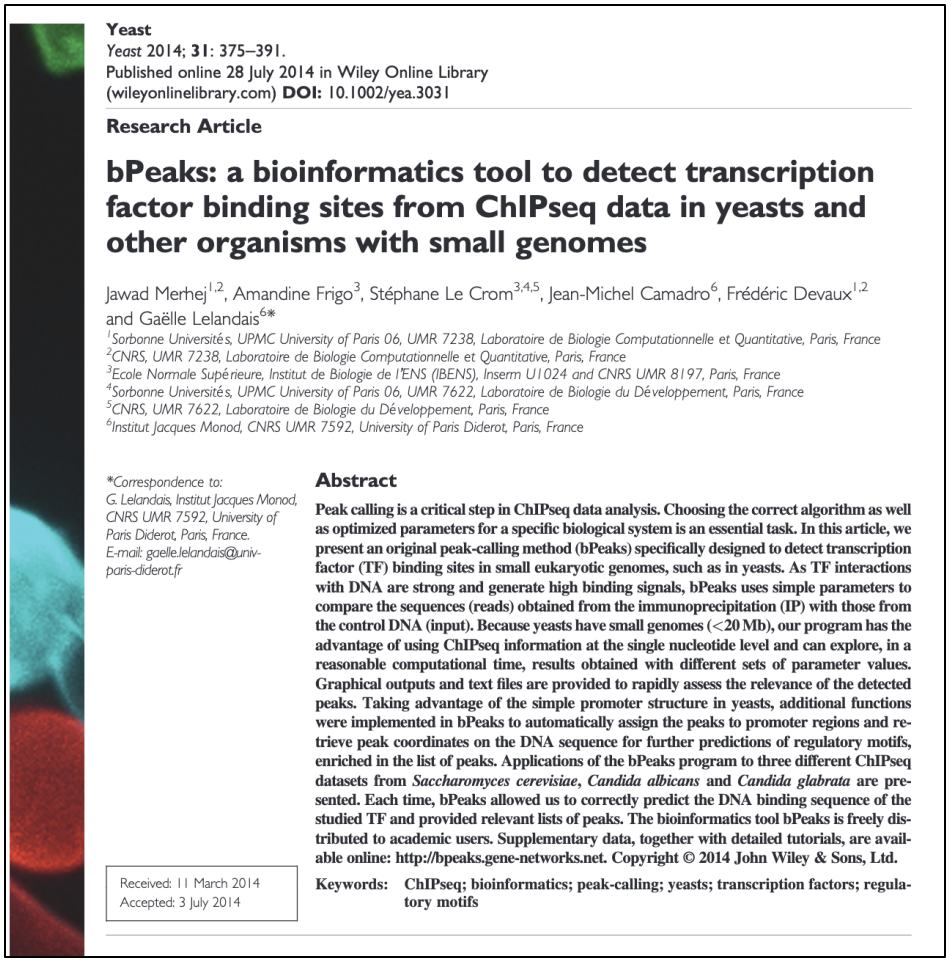
Abstract

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome, allowing for more robust predictions. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, and is freely available.





My favorite method is...



Yeast
Yeast 2014; **31**: 375–391.
Published online 28 July 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/yea.3031

Research Article

bPeaks: a bioinformatics tool to detect transcription factor binding sites from ChIPseq data in yeasts and other organisms with small genomes

Jawad Merhej[1,2], Amandine Frigo[3], Stéphane Le Crom[3,4,5], Jean-Michel Camadro[6], Frédéric Devaux[1,2] and Gaëlle Lelandais[6]*

[1]*Sorbonne Universités, UPMC University of Paris 06, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France*
[2]*CNRS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France*
[3]*Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), Inserm U1024 and CNRS UMR 8197, Paris, France*
[4]*Sorbonne Universités, UPMC University of Paris 06, UMR 7622, Laboratoire de Biologie du Développement, Paris, France*
[5]*CNRS, UMR 7622, Laboratoire de Biologie du Développement, Paris, France*
[6]*Institut Jacques Monod, CNRS UMR 7592, University of Paris Diderot, Paris, France*

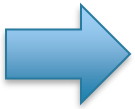
*Correspondence to:
G. Lelandais, Institut Jacques Monod, CNRS UMR 7592, University of Paris Diderot, Paris, France.
E-mail: gaelle.lelandais@univ-paris-diderot.fr

Abstract

Peak calling is a critical step in ChIPseq data analysis. Choosing the correct algorithm as well as optimized parameters for a specific biological system is an essential task. In this article, we present an original peak-calling method (bPeaks) specifically designed to detect transcription factor (TF) binding sites in small eukaryotic genomes, such as in yeasts. As TF interactions with DNA are strong and generate high binding signals, bPeaks uses simple parameters to compare the sequences (reads) obtained from the immunoprecipitation (IP) with those from the control DNA (input). Because yeasts have small genomes (<20 Mb), our program has the advantage of using ChIPseq information at the single nucleotide level and can explore, in a reasonable computational time, results obtained with different sets of parameter values. Graphical outputs and text files are provided to rapidly assess the relevance of the detected peaks. Taking advantage of the simple promoter structure in yeasts, additional functions were implemented in bPeaks to automatically assign the peaks to promoter regions and retrieve peak coordinates on the DNA sequence for further predictions of regulatory motifs, enriched in the list of peaks. Applications of the bPeaks program to three different ChIPseq datasets from *Saccharomyces cerevisiae*, *Candida albicans* and *Candida glabrata* are presented. Each time, bPeaks allowed us to correctly predict the DNA binding sequence of the studied TF and provided relevant lists of peaks. The bioinformatics tool bPeaks is freely distributed to academic users. Supplementary data, together with detailed tutorials, are available online: http://bpeaks.gene-networks.net.

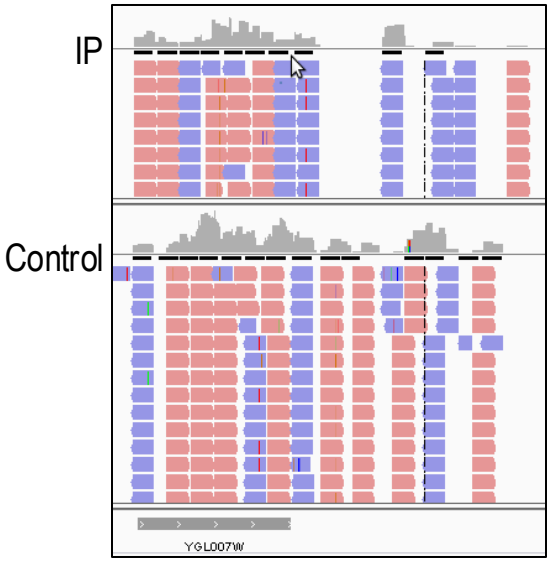
Received: 11 March 2014
Accepted: 3 July 2014

Keywords: **ChIPseq; bioinformatics; peak-calling; yeasts; transcription factors; regulatory motifs**

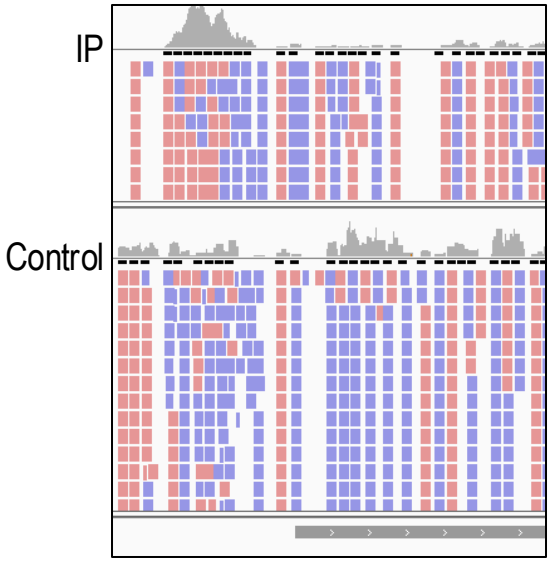


(Impossible) challenge faced by peak calling programs

Bad peak ☹



Good peak ☺



distinguishing 'real' peaks (interaction between TF and DNA)

from 'artefact' peaks (other peaks)

Computional biology = Patience

(Impossible) challenge faced by peak calling programs

distinguishing ‘real’ peaks (interaction between TF and DNA)

from ‘artefact’ peaks (other peaks)

Computional biology = Patience

Multiple information to be used to validate peak calling results