Abordagens Ómicas em Biomedicina e Biotecnologia

Ciências ULisboa

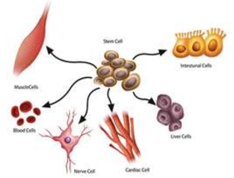
ChIP-seq 101

Paulo Matos

2020/2021

Regulation of Gene Expression…

- One genome ➔ multiple cell types!

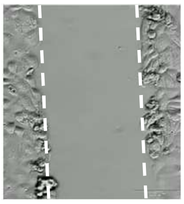


Precise regulation of gene expression is determinant for the stimulus-cued and temporally and spatially coordinated differentiation and function of different cell-types…

https://www.cryo-cell.com/CryoCell/media/CryoCell/Client-images/stem-cell-differentiation.png

---

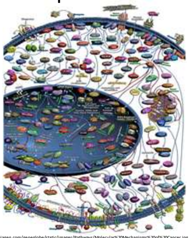
Regulation of Gene Expression…



Epithelial cell monolayer

Regulation of Gene Expression…

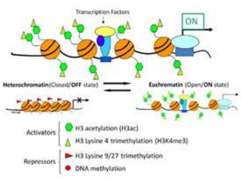
- Different stimuli are temporally and spatially integrated through complex signal transduction pathways and relayed to the nucleus…



https://www.qiagen.com/geneglobe/static/images/Pathways/Molecular%20Mechanisms%20of%20Cancer.jpg

---

Transcriptional regulation in eukaryotes

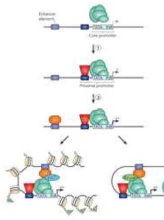


Baldwin (2013) Front Physiol 4:284

Levels of transcriptional regulation:

- 3D structure of chromatin in the nucleus - Chromatin compaction by histone proteins and their post-translational modifications;
- Reversible modifications of DNA (e.g., methylation of promoters);
- Transcription factor activation and binding to DNA…

Transcription factors in eukaryotes

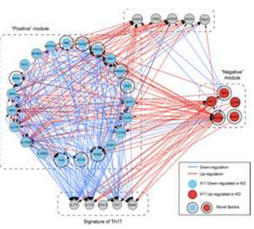


Farnham, 2009. Nature reviews. Genetics, 10(9):605-16.

Transcription factors (TFs) are proteins that bind specific DNA sequences to control the transcription of genes.

- General TFs bind to core promoter regions.
- Site-specific TFs bind at the proximal promoter or distal enhancer or repressor regions.
- TFs often work in protein complexes, integrating the outputs of several signaling pathways…

Identification of TF binding sites (TFBS)



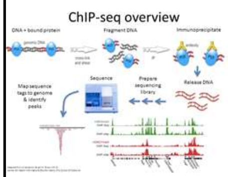
Yosef et al., 2013. Nature 496:461–468.

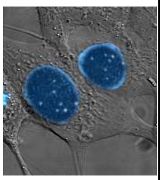
To understand how different conditions or signals influence gene expression and how particular regulatory molecules or events are involved in the observed responses, we need to know:

- **All of the sites in the genome** to which the **transcription regulators** of interest bind, under those conditions.
- Possibly **correlate** this information with **chromatin status**, and transcript **expression levels**.

ChIP-seq

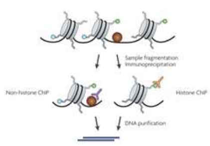
Chromatin **I**mmuno**P**recipitation-next generation **Seq**uencing





ChIP-Seq principle

Find all regions in the genome bound by a specific transcription factor (or bearing a specific histone modification) in a certain experimental setting (a particular cell type, a specific stimulus, a drug treatment, etc…).

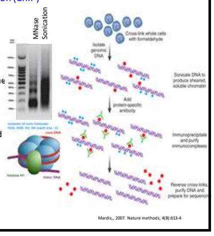


Park, 2009. Nature reviews. Genetics, 10(10):669(80.

ChIP-Seq method

Step 1: Chromatin Immunoprecipitation (ChIP)

- Covalent cross-links between proteins and DNA are formed *in vivo* (formaldehyde is the most common cross-linker…)
- Sheared chromatin (by sonication or using micrococcal nuclease) produce the protein-DNA fragment complexes (usually 150-300bp…)
- An antibody specific to the protein (or protein modification) of interest is used to selectively immunoprecipitate that were cross-linked (**highly specific ChIP-grade Abs are mandatory**…).
- The protein-DNA complexes are de-cross-linked and the IP-enriched DNA fraction is purified.



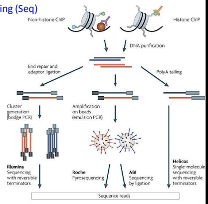
Mardis,, 2007. Nature methods, 4(8):613-4

ChIP-Seq method

Step 2: high-throughput DNA sequencing (Seq)

- DNA sequencing libraries are generated by clonal amplification by PCR in vitro.
- The amplified DNA templates are sequenced simultaneously in a massively parallel fashion producing millions of "**reads**".

Note: While the basic workflow is common to most NGS platforms, each utilizes slightly different strategy…



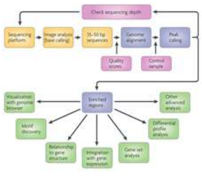
Park, 2009. Nature reviews. Genetics, 10(10):669(80.

ChIP-Seq method

Step 3: bioinformatic analysis of ChIP-Seq data

General workflow of ChIP-seq analysis:

I. Quality control of raw sequencing data.
II. Mapping reads to the reference genome.
III. Peak calling to determine binding locations.
IV. Functional analyses
   - Associate peaks to genes
   - GO analysis of targeted genes
   - Identification of TF binding motifs
   - ...



Park, 2009. Nature reviews. Genetics, 10(10):669(80.

ChIP-Seq raw data

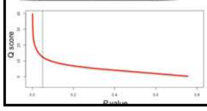
Illumina HiSeq can produce >50.000.000 reads of 36-100 bp/ChIP-seq experiment*

*HiSeq® 2500 Sequencing System manual

- ChIP-seq raw data consists of sequencing reads.
- Raw sequence reads from Illumina are reported in **FASTQ format**.
- FASTQ format is a text-based format for storing both a nucleotide sequence and its corresponding **Quality Scores**.
- Illumina quality score (aka, Phred score) of a base is calculated as :

$$Q=-10 \log_{10} p$$

where ***p*** is the estimated probability of a base being wrong.

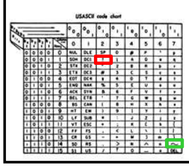


Q values can encode Phred scores from 0 to 93, **where $p < 0.05$ when $Q > 13$…**

ChIP-Seq raw data

A FASTQ file normally uses four lines per each read

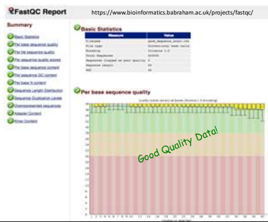
```
@SRS_2neu_2_1_1520_1070
AAAAAGTCCCCCGCGTTCCCAGCTTGGGGTGGTTTG
+
E@8D8GEFGDGGGADGGGDGBGD8BBBB?7@@=FAD
```



- **Line 1** begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- **Line 2** is the raw sequence letters.
- **Line 3** begins with a '+' character and is optionally followed by the same sequence identifier again (and any other description).
- **Line 4** correspond to ASCII encoded base quality values for the sequence in Line 2 [ASCII(Q+33), i.e., "!" is the lowest and "~" the highest].

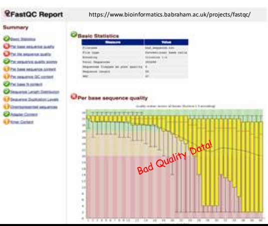
Quality control of ChIP-seq reads

Sequence read quality can be assessed using the FastQC tool



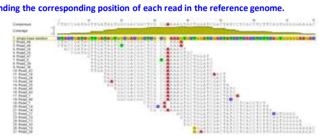
Quality control of ChIP-seq reads

Sequence read quality can be assessed using the FastQC tool



Mapping reads to the genome

Finding the corresponding position of each read in the reference genome.



- The most computationally demanding step of the analysis - needs high performance computers.
- Requires ultrafast and memory-efficient algorithms to compress genomic data (Borrows-wheeler transform) while still permitting fast substring (read) queries.
- Read alignment must also **allow for a small number of gaps and mismatches** (sequencing errors and variations in sample genome)

Mapping reads to the genome

Bowtie2 is a commonly used mapping tool for ChIP-Seq data

Bowtie 2

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

The **Bowtie2** mapping tool is recommended by the **Encode** project for ChIP-seq pipelines

ENCODE: the **Enc**yclopedia **O**f **D**NA **E**lements is a project launched by the NIH's Human Genome Research Institute in 2003 that **aims to identify all functional elements in the human genome sequence.**

Mapping reads to the genome

Bowtie2 is a commonly used mapping tool for ChIP-Seq data

Bowtie 2
Fast and sensitive read alignment

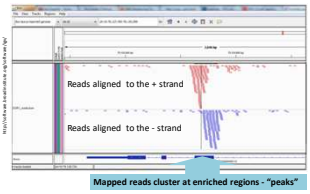
Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Bowtie output files are in the SAM (Sequence Alignment/Map) format
- A generic format for storing large nucleotide sequence alignments.
- The related **BAM format** is a compressed binary representation of a SAM files.

Field	Description
QNAME	Query template (read) NAME
FLAG	A bit value summarizing several alignment control parameters
RNAME	References sequence NAME
POS	1- based leftmost mapping POSition
MAPQ	MAPping Quality
CIGAR	Mismatch report string
RNEXT	Ref. name of the mate/next read
PNEXT	Position of the mate/next read
TLEN	observed Template LENgth
SEQ	segment SEQuence
QUAL	ASCII of Phred-scaled base QUALity+33

Mapping reads to the genome

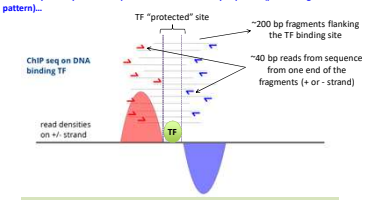
Mapped ChIP-seq data can be visualized as tracks in genome browser tools [e.g., the Integrated Genome Browser (IGV) or the UCSC Genome Browser] together with other annotations (e.g., genes in the vicinity)...



Mapped reads cluster at enriched regions - "peaks"

Peak Calling

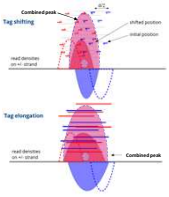
Strand asymmetry in ChIP-seq read densities is actually expected (peak recognition pattern)...



+/- strand read "peaks" do not represent the true location of the binding site

Peak Calling

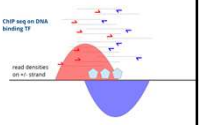
- ✓ **The input signal for peak calling is the number of mapped reads at each genomic position.**
- ✓ **Reads mapping to the forward and reverse strand have to be combined to increase the signal and peak resolution.**

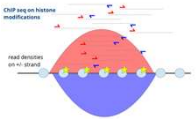


- **Tag Shifting:** reads are *shifted* by *d*/2 (i.e. towards the middle of the IP fragment), where *d* represent the estimated mean fragment length.
- **Tag enlongation:** Each tag is computationally extended in 3' to a total length of *d*.

Peak Calling

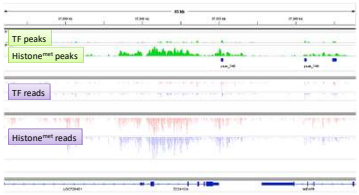
- ✓ Strand asymmetry is blurred by the binding of several TFs as complexes...



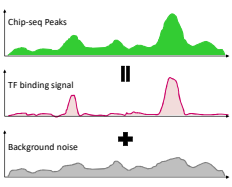


...and is virtually lost in datasets of ChIP for histone modifications.

For example...



ChIP = target enrichment...



Potential sources of "noise" in ChIP-seq experiments:
- DNA-shearing is not uniform across genome, which results in more reads in open chromatin regions.
- PCR amplification bias (GC content) in library preparation.
- Repetitive regions might appear enriched due to underestimated repeat copies in the reference genome.

How do we estimate the noise?

"Data don't make any sense,
we will have to resort to statistics."

Poisson model for read counts

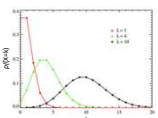
If we assume that the **"noise"reads** within a given genomic window have a **random distribution** then we can use a theoretical distribution to estimate the **probability of the observed number of reads being above what is expected (i.e., the "noise").**

The **Poisson distribution** is considered adequate since it models the number of random events (reads) in a fixed interval (genomic window) (*Note - some authors prefer the binomial or the negative binomial*):

$$P(X=k)=e^{-k}\frac{\lambda_c^k}{k!}$$

P = the probability that ***k*** reads map to a given genomic region.

λ = mean number of reads in a genomic window of a certain size.



Certain algorithms use a **uniform λ value** estimated from annotated whole genome data. But....

Modelling noise levels

Noise is **not uniform** (chromatin conformation, technical biases, repetitive regions...etc)

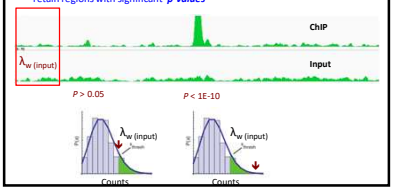
Solution: sequence **Input DNA to use as reference** (control)!

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

Defining "significant peaks"

Slide a window across the genome. At each location:

- Evaluate the enrichement of the ChIP signal
- Caculate **λ** from input and ***p*** based on the poison distribution
- retain regions with significant ***p-values***



MACS (Model-based Analysis of ChIP-Seq)

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

MACS is the pipeline recommended by the **Encode** project for ChIP-seq data analysis

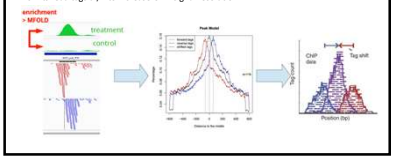
ENCODE: the **Enc**yclopedia **O**f **DN**A Elements is a project launched by the NIH's Human Genome Research Institute in 2003 that **aims to identify all functional elements in the human genome sequence.**

MACS (Model-based Analysis of ChIP-Seq)

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

1 : Modeling the shift size of ChIP-Seq tags (estimating *d*)

- Slides a window of 2x the user defined sonication size (BANDWITH)
- Detects regions with high read enrichment in ChIP vs. input (MFOLD - default 10-30 fold)
- Plots average +/- strand read densities → estimates ***d'*** (i.e, mean fragment size)
- Shifts read tags *d'*/2 to increase ChIP signal resolution...



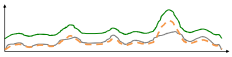
MACS (Model-based Analysis of ChIP-Seq)

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

2 : "Significant Peak" detection

I. **Normalize datasets:**

- Datasets generally do not have the same sequencing depth
- ChIP (=signal + noise) and input (=noise)



Input = N reads
ChIP-seq dataset = M > N **reads**
ChIP scaled by library size :
M' = M*(N/M)

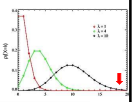
MACS (Model-based Analysis of ChIP-Seq)

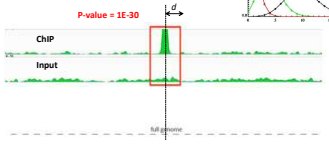
Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

2 : "Significant Peak" detection

II. **Identify local noise parameter**

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





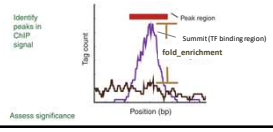
MACS (Model-based Analysis of ChIP-Seq)

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

2 : "Significant Peak" detection

III. **Calling peaks and scoring**

- Candidate peaks with **p-values below** a user-defined threshold (PVALUE - default 1E-05) are called (**Peak regions**);
- The ratio between the ChIP-Seq tag count and λ_{local} is reported as **fold_enrichment.**
- The location with the **highest fragment pileup (summit)** is predicted as the TF binding location.



MACS (Model-based Analysis of ChIP-Seq)

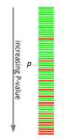
Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

3 : Estimation of False Discovery Rate (FDR)

- The FDR is an estimation of the proportion of incorrectly identified peaks among those that are found to be of a certain significance (p-value).

- **How MACS does it?**
  - Swaps "ChIP" and "input" datasets and calculates p-values for **"negative peaks".**
  - Ranks p-values from both **Positive** (ChIP vs input) and **Negative** (input vs. ChIP) peaks.
  - Calculates FDR for a peak with a given ***p*** as:

$$FDR(p) = \frac{\#\ \textbf{negative}\ \text{peaks with Pval} < p}{\#\ \textbf{positive}\ \text{peaks with Pval} < p}$$



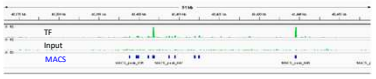
MACS (Model-based Analysis of ChIP-Seq)

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

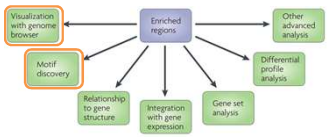
4 : Output a list of peaks coordinates in BED format

- **BED** (Browser Extensible Data) is a UCSC Genome Browser format that provides a flexible way to define the data lines that are displayed in an annotation track.

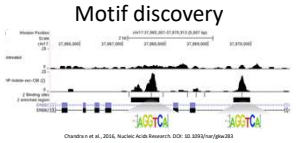
chr	start	end	length	summit	tags	-10*log10(pvalue)	fold_enrichment	FDR(%)
chr1	8250	8671	422	206	66	140.84	11.68	0.51
chr1	36382	36984	603	405	66	315.23	27.05	0.24
chr1	71082	70475	394	273	16	65.40	3.74	1.13
chr1	389664	390025	362	118	13	62.30	7.57	1.19
chr1	414983	416588	606	279	41	234.40	21.86	0.35
chr1	449089	450054	966	665	90	368.81	20.10	0.19
chr1	450217	450908	692	245	77	349.57	18.81	0.20
chr1	512406	513125	720	336	66	470.54	36.90	0.17
chr1	513304	514143	840	130	60	369.72	27.93	0.19



Downstream analyses

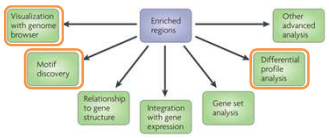


Motif discovery

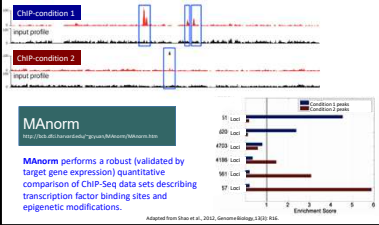


MEME-ChIP performs comprehensive motif analysis (including motif discovery, motif enrichment analysis and clustering) on large nucleotide datasets (50MB maximum) such as those identified by ChIP-seq.

Downstream analyses

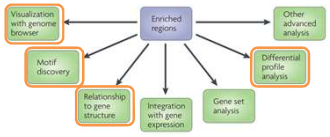


Differential profile analysis



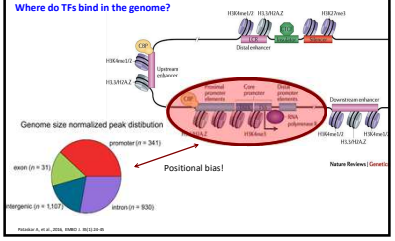
MAnorm performs a robust (validated by target gene expression) quantitative comparison of ChIP-Seq data sets describing transcription factor binding sites and epigenetic modifications.

Downstream analyses



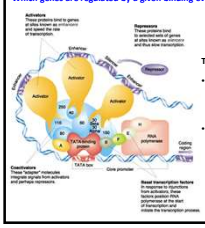
Relationship to gene structure

Where do TFs bind in the genome?



Association of peaks to genes

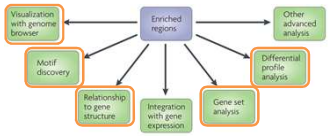
Which genes are regulated by a given binding event?



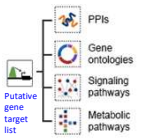
Two commonly used approaches:

- Associate the TF binding site (TFBS) to all transcriptional start sites (TSS) within a fixed window (e.g. 20kb).
- Associate the TF binding site (TFBS) to the nearest TSS.

Downstream analyses



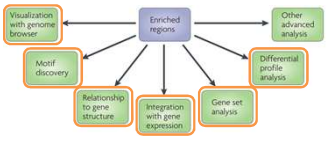
Gene set analysis



Enriched functions among the genes associated to (i.e. "close to") the predicted TFBS can be explored with:

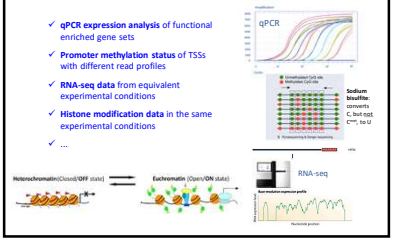
- GO (Gene Ontology) - **DAVID** (Huang et al., NAR 2009)
- **KEGG** or **REACTOME** (pathways)
- **BioGRID** (protein and genetic interaction)
- etc...

Downstream analyses



Integration with gene expression data

- ✓ **qPCR expression analysis** of functional enriched gene sets
- ✓ **Promoter methylation status** of TSSs with different read profiles
- ✓ **RNA-seq data** from equivalent experimental conditions
- ✓ **Histone modification data** in the same experimental conditions
- ✓ ...



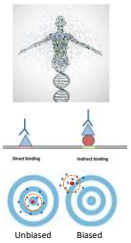
Advantages and disadvantages of ChIP-seq

Advantages:

- Genome-wide binding sites.
- High resolution (~200 bp)
- Relatively cheap and scalable (*ENCODE project performed hundreds of ChIP-seq experiment for many TFs and histone marks in many tissues, conditions and organisms*).

Disadvantages:

- Requires large cell populations (millions)
  (although single-cell ChIP-Seq approaches are being developed)
- Depends on highly specific antibodies
- Can not distinguish between direct and indirect TF binding
- Finding enriched binding regions (peaks) relies heavily on statistical methods - bias towards genomic amplifications (cancer cells), repetitive regions, and **high affinity binding sites...**



Suggested reading

- ✓ Nakatoa, R. and Sakatab, T. (2020) Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods, in press; DOI: 10.1016/j.ymeth.2020.03.005
- ✓ Pavesi, G. (2017) ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks. Adv Biochem Eng Biotechnol., 160:1-14.
- ✓ Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature reviews. Genetics, 13:840-52.