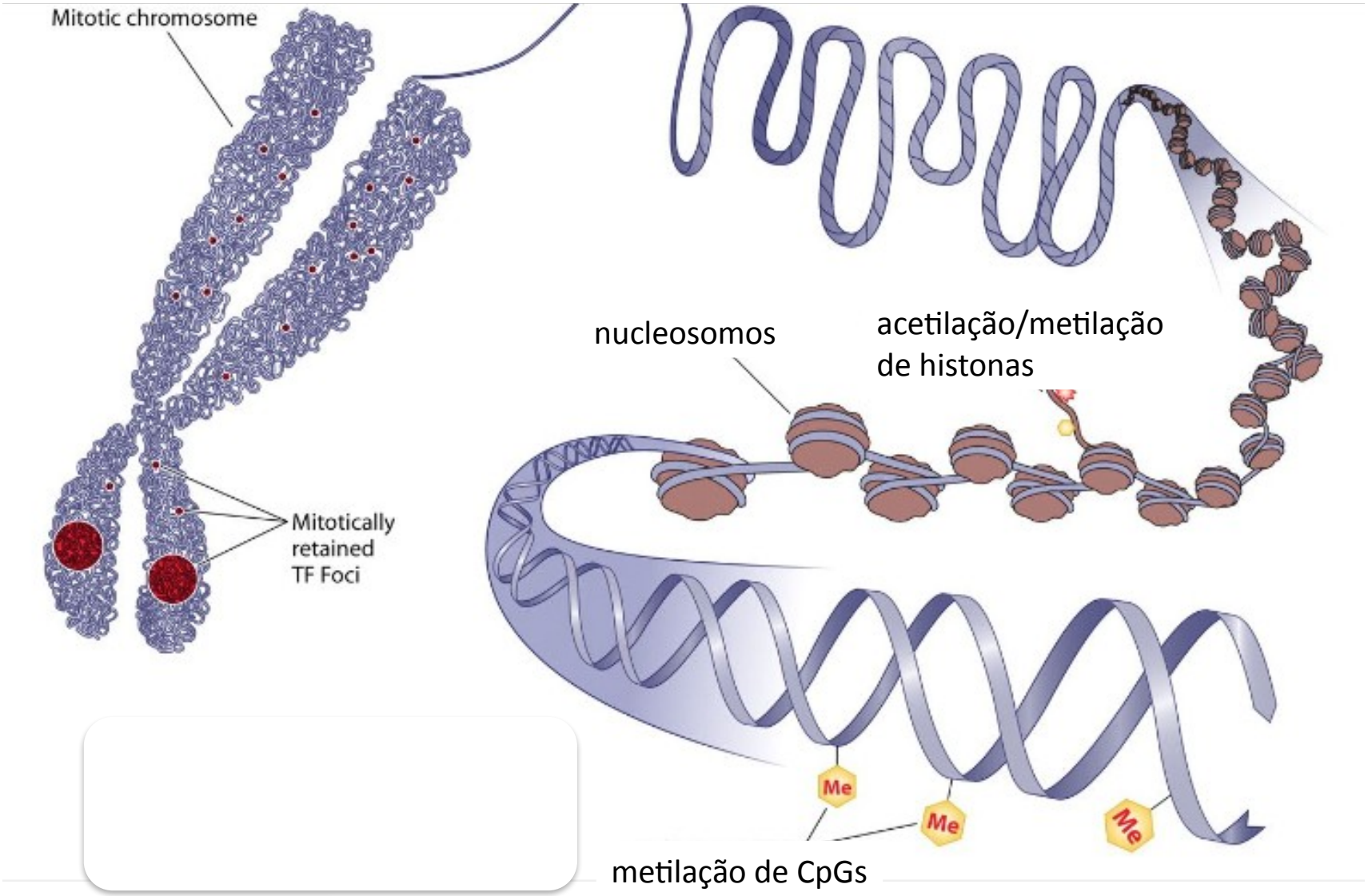
Biologia Molecular Computacional
IBI5035/QBQ2507 - 2023

Análise global de elementos regulatórios da expressão gênica

análise de dados de ChIP-Seq

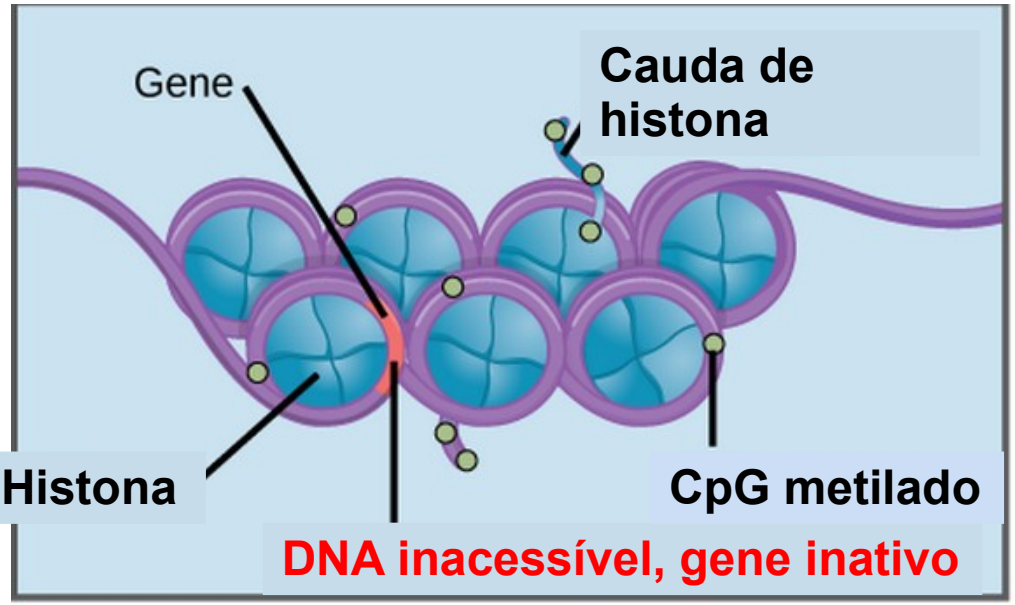
Eduardo Moraes Rego Reis
Instituto de Química - USP

Modificações de histonas e **metilação do DNA** são mecanismos importantes para a **regulação** da expressão gênica



adaptado de Zaidi SK et al.. Mol Cell Biol. 2010

Modificações covalentes reversíveis em proteinas histonas definem o estados de ativação ou repressão gênica



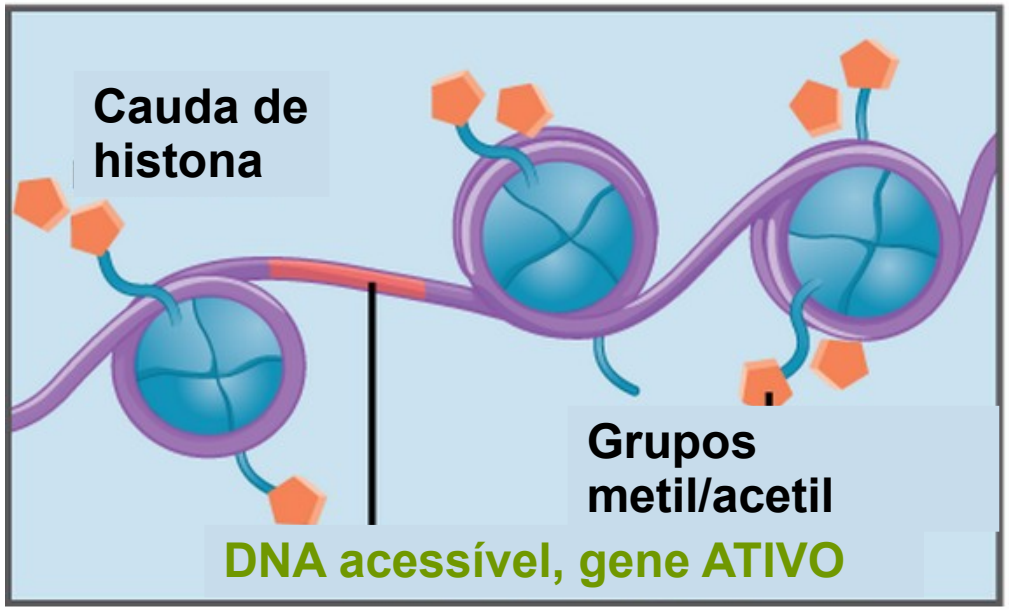
Histonas deacetiladas/demetiladas, CpG metilados

⇩

Nucleosomos altamente empacotados.

⇩

A maquinaria de transcrição não consegue acessar o DNA.



Histonas acetiladas/metiladas, CpGs demetilados

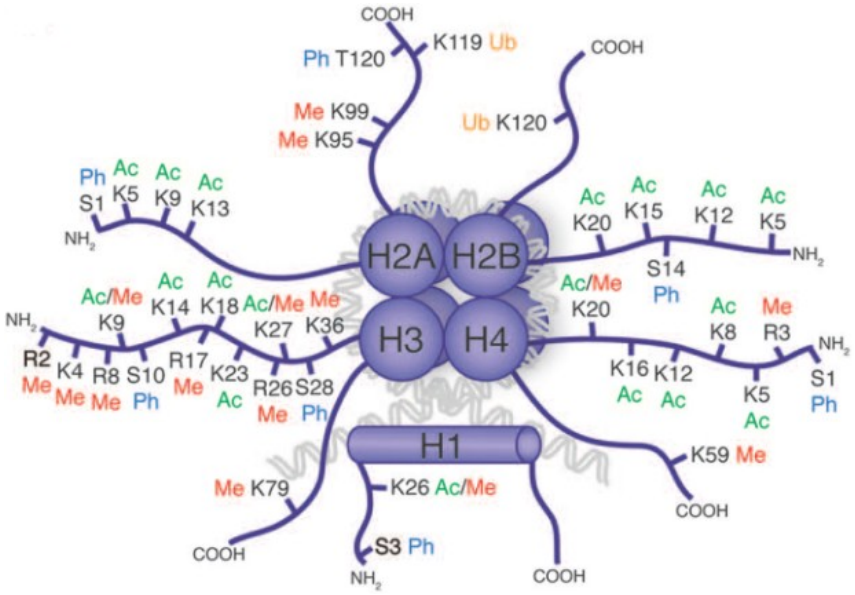
⇩

Diminuição do empacotamento dos nucleosomos.

⇩

DNA se torna mais acessível a ligação de fatores de transcrição ativadores

Modificações de histonas controlam a compactação do DNA e a expressão gênica



		Papel na transcrição gênica	Resíduos modificados na histona
Acetilação	$COCH_3$	ativação	H3 (K9, K14, K18, K56) H4 (K5, K8, K12, K16) H2A/H2B (K6, K7, K16, K17)
Fosforilação	PO_3	ativação	H3 (S10)
Metilação	CH_3	ativação	H3 (K4, K36, K79)
		repressão	H4 (K20)
Ubiquitinação	156 aa	ativação	H2B (K123)
		repressão	H2A (K119)
Sumoilação	101 aa	repressão	H3 (?) H4 (K5, K8, K12, K16) H2A (K126) H2B (K6, K7, K16, K17)

- Combinações de diferentes marcas tem efeito combinatorial na cromatina
- Código das histonas ainda não é totalmente conhecido

Promotor

H3K4me3
H3K27ac
H3K9ac

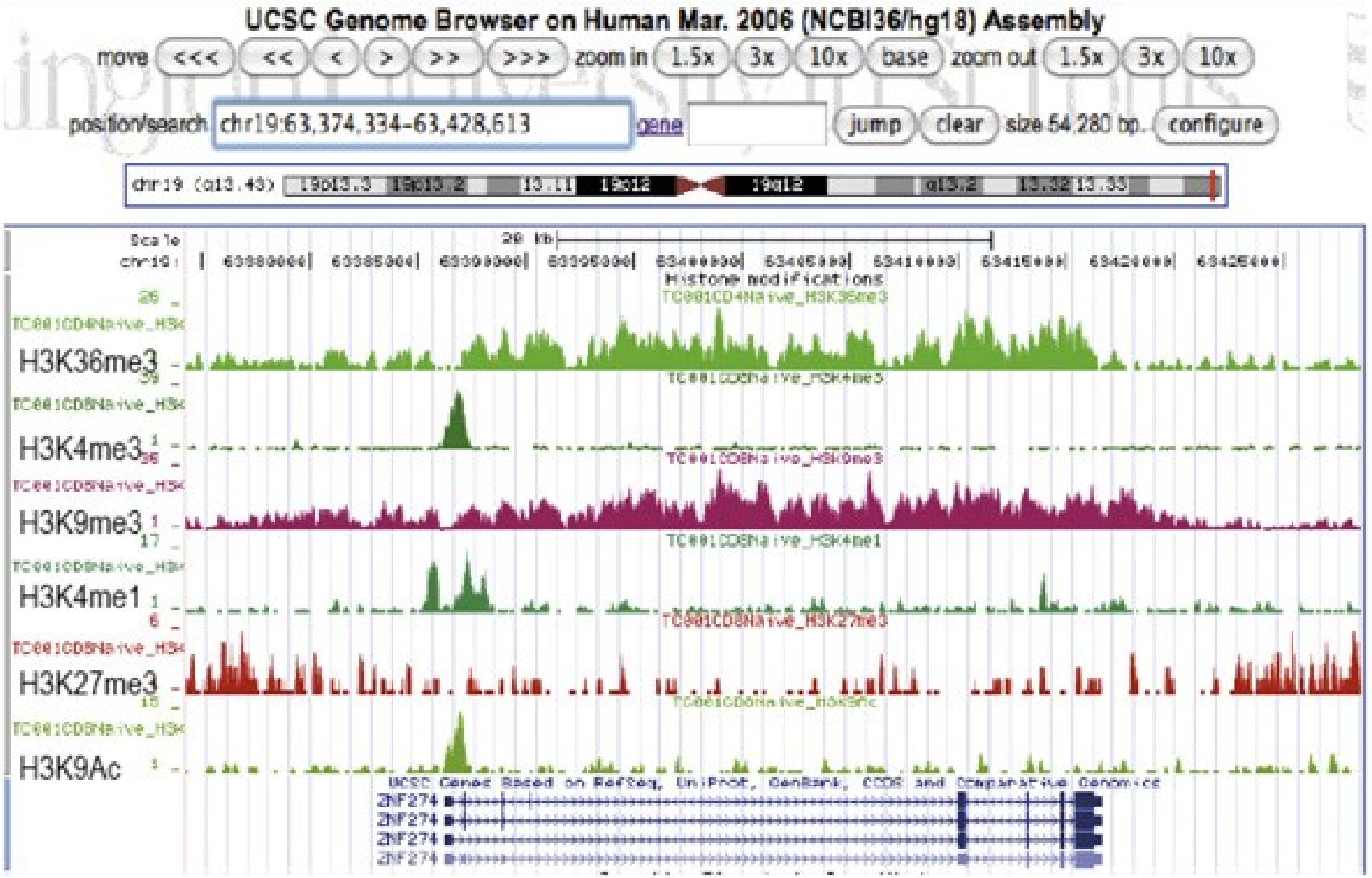
H3K27me3

enhancers proximais e distais

H3K27ac

Mapas de estados da cromatina baseados na localização de marcas de histonas informa sobre a atividae dos genes

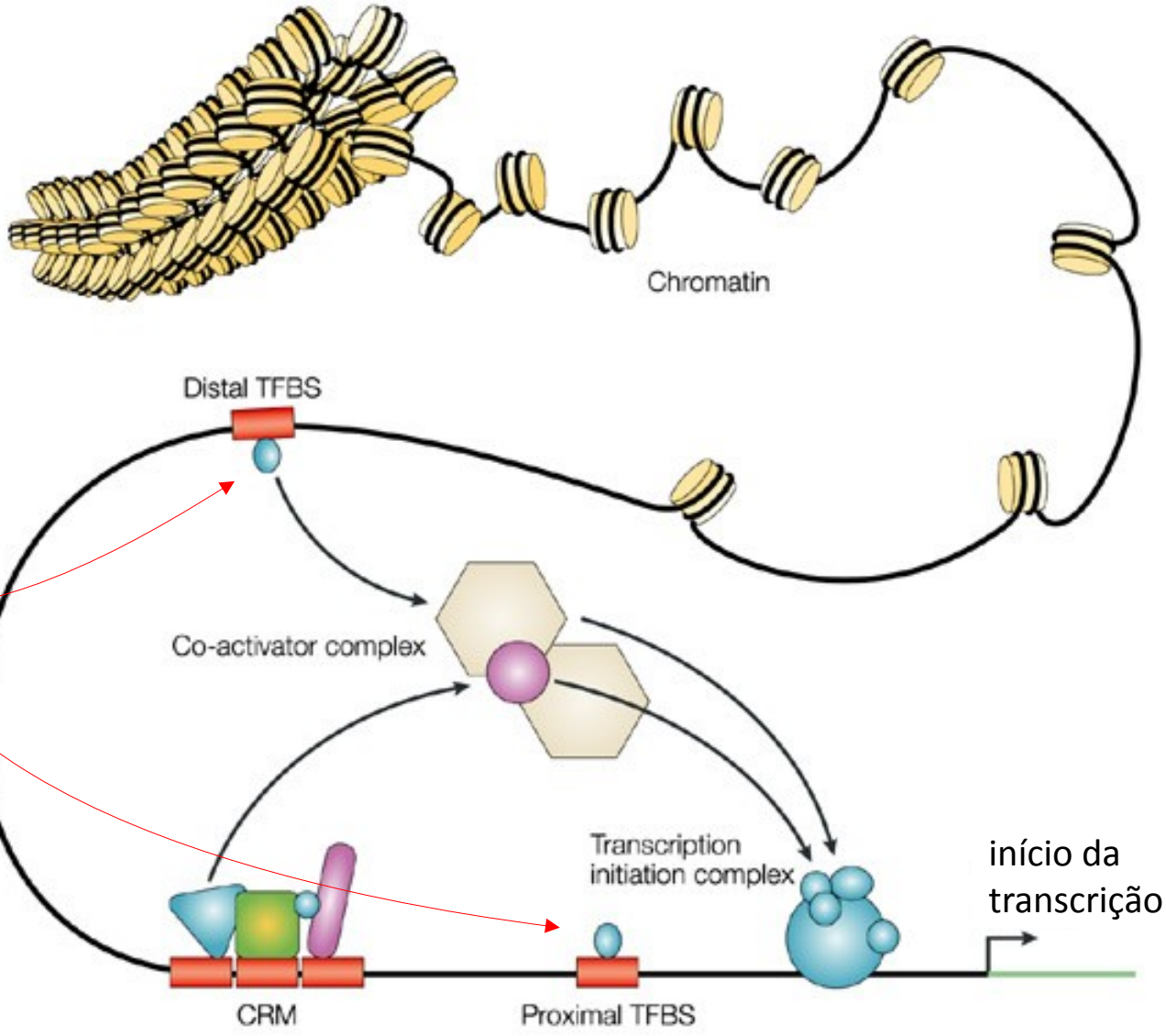
visualização de dados de ChIP-seq em browsers genômicos



Fatores de transcrição se ligam ao DNA e ativam ou reprimem a transcrição

TFBS:
sítios de ligação de **fatores de transcrição**

Ativadores da expressão gênica



Wasserman e Sandelin, ANat Rev Genet 5, 276–287 (2004)

ChIP-seq:

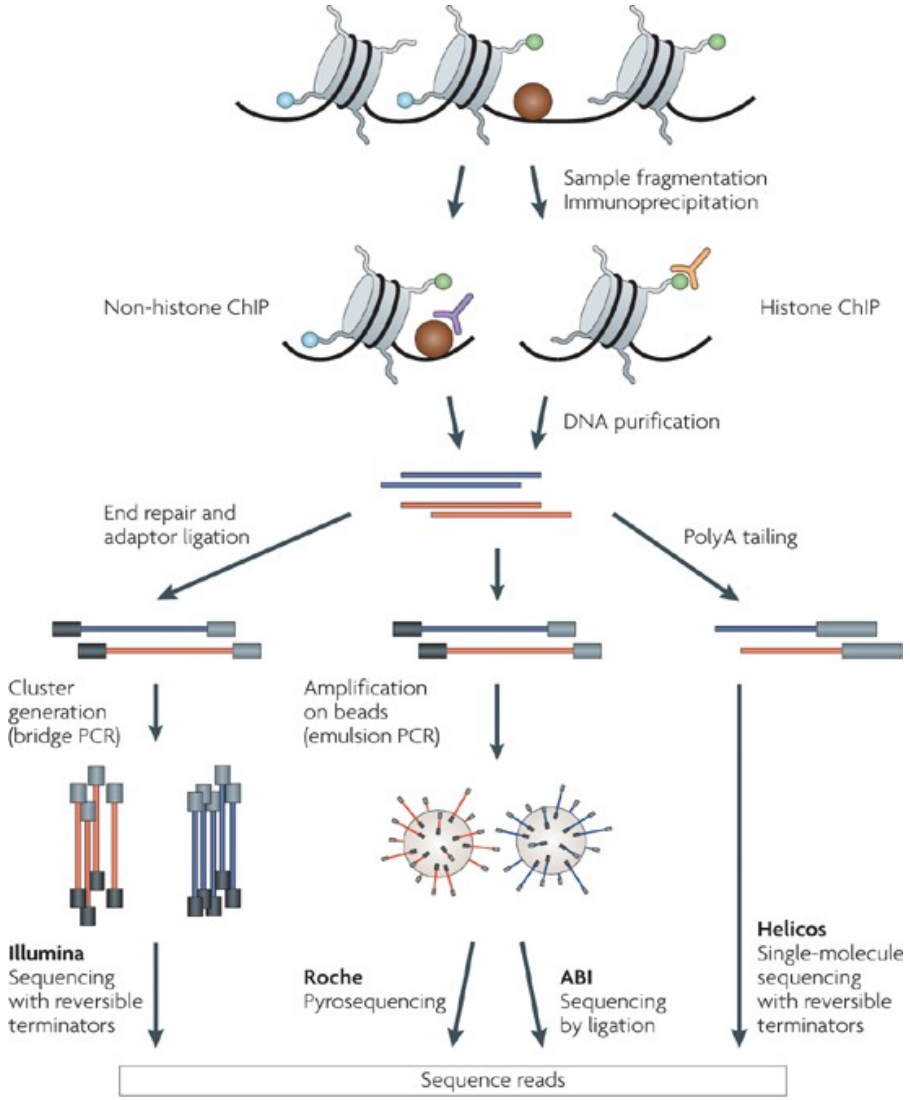
Imunoprecipitação de proteínas associadas a regiões específicas do DNA seguido de sequenciamento

São utilizados anticorpos que reconhecem especificamente a proteína de interesse associada ao DNA.

A cromatina é fragmentada e o DNA associado a proteina-alvo é recuperado por imunoprecipitação e analisado por NGS.

Permite identificar globalmente os sítios de ligação no DNA de :

- acetilação/metilação de histonas
- Fatores de transcrição

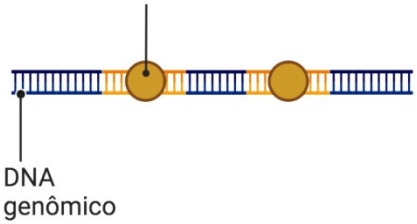


Nature Reviews Genetics 10, 669-680 (October 2009)

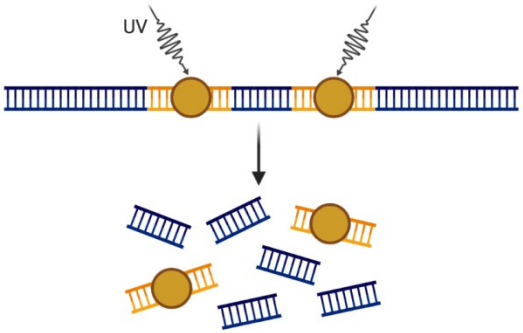
Etapas de um experimento de ChIP-seq

Identificar locais no genoma contendo marcas epigenéticas ou ocupados por proteínas regulatórias da cromatina

- Modificações de histonas
- Fatores de transcrição

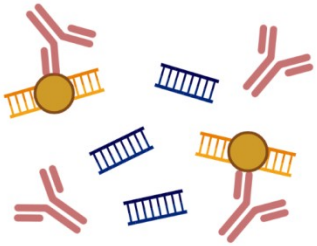


1 Crosslinking e fragmentação do DNA

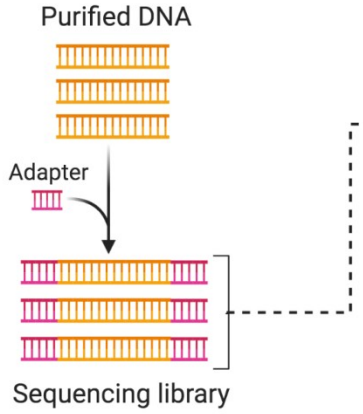


2 Immunoprecipitação da proteína e do DNA ligado

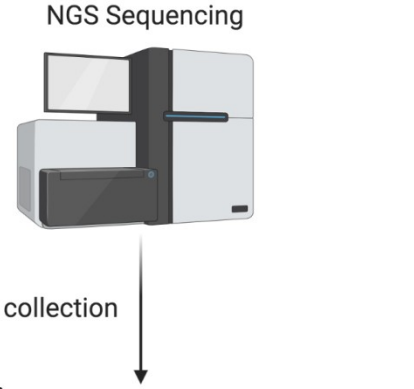
anticorpo específico contra a proteína-alvo



3 purificação do DNA e ligação de adaptadores para NGS



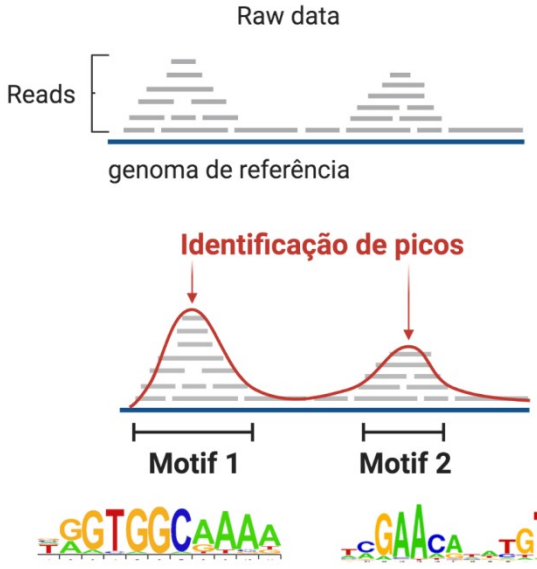
4 Sequenciamento NGS



Reads:

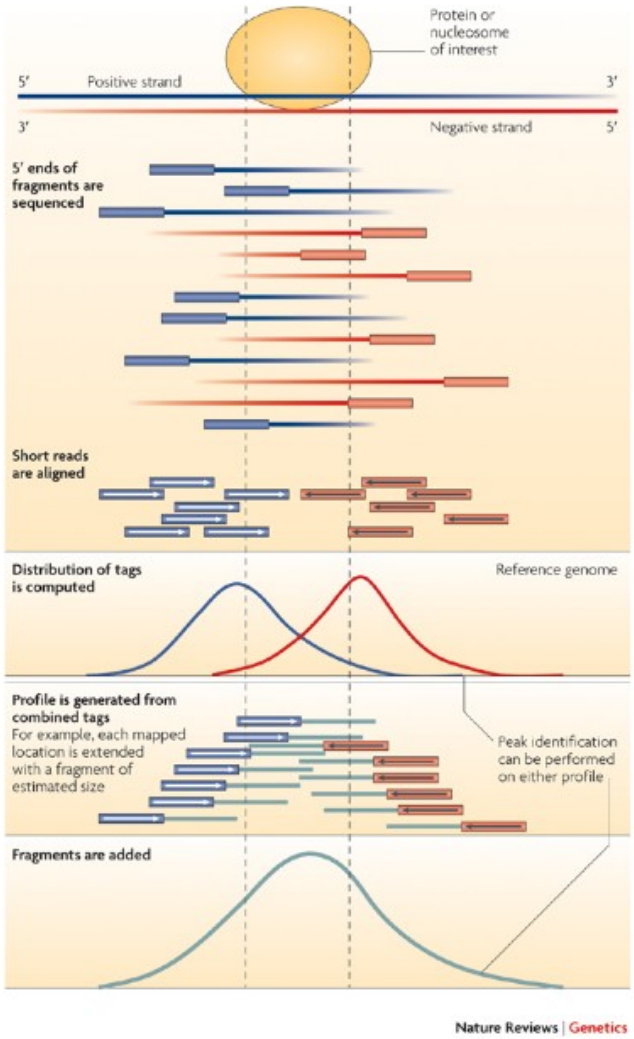
....GTTCCTTCAGCATTTGCAGGTACCGT....

5 Identificação de sítios de ligação no genoma e de motivos no DNA



Exemplo de pipeline computacional para a análise de dados de ChIP-seq

- DNA associado a proteina-alvo co-imunoprecipitado com anticorpo (ChIP)
- DNA total (input)



Park, P. J., *ChIP-seq: advantages and challenges a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

Passo 1: verificar a qualidade dos reads (FastQC)

Passo 2: Alinhar os reads no genoma (BWA, Bowtie)

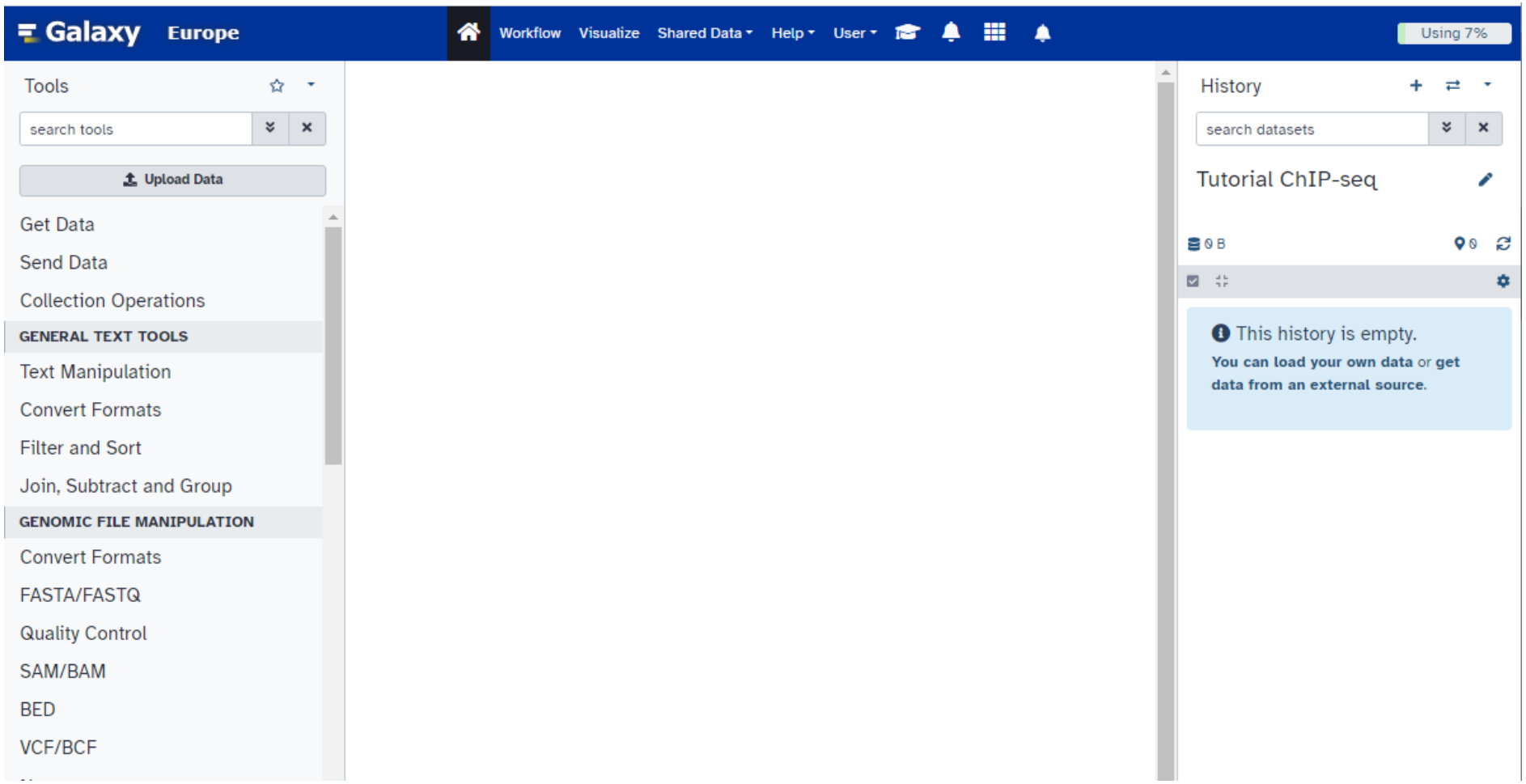
Passo 3: Identificação de picos (MACS2)

Passo 4: anotação de picos
- Identificação de genes sobrepostos
- Identificação de picos comuns/exclusivos entre duas ou mais condições

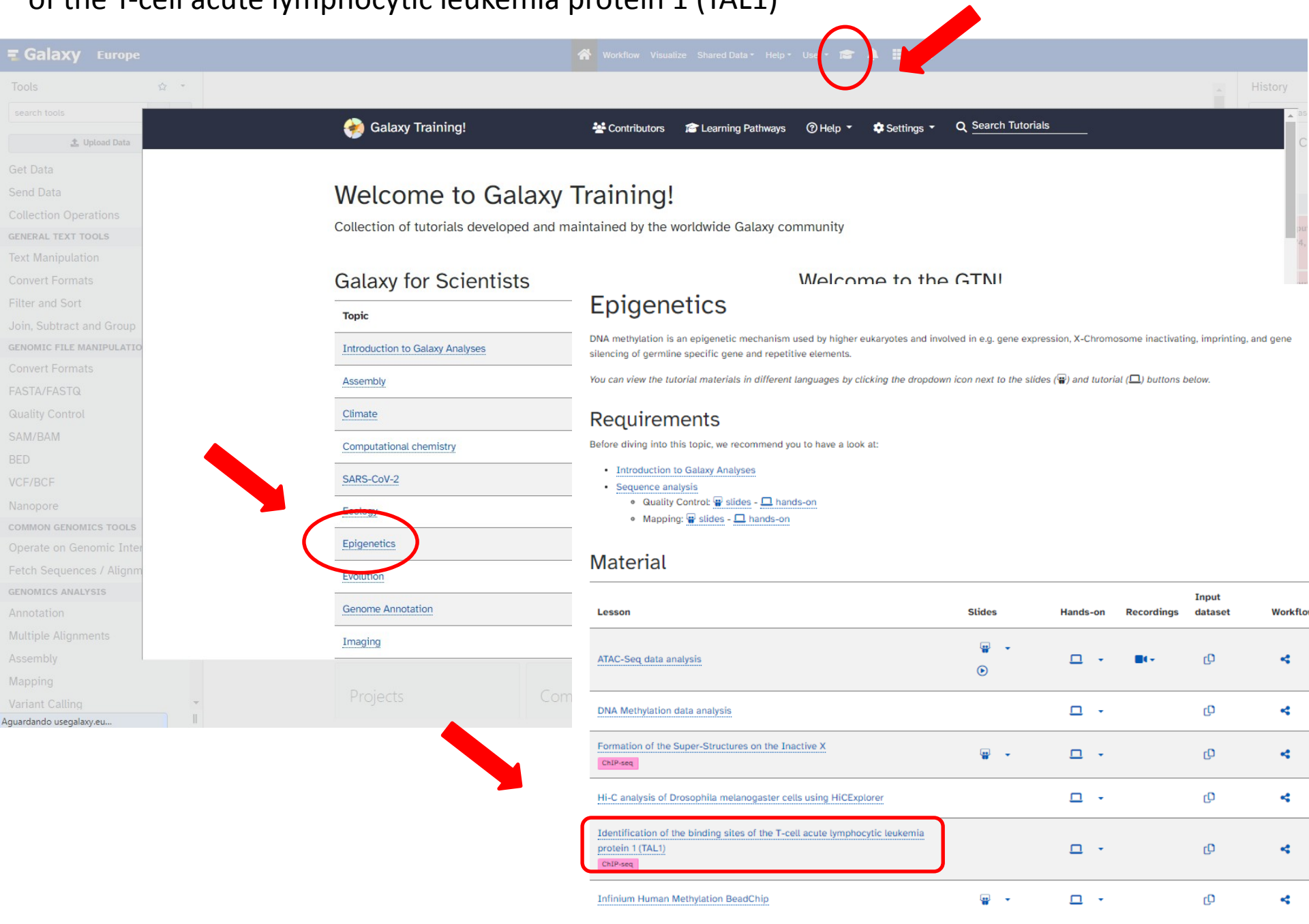
- Uso do DNA input como referência aumenta a acurácia pois reduz o número de falsos positivos (regiões de cromatina aberta tendem a ser mais detectadas)

Tutorial – ChIP Seq (Galaxy)

No Galaxy Europe,
criar um “history” e renomear como “tutorial ChIP-seq”



No modo “tutorial” do Galaxy, navegar até o tutorial “Identification of the binding sites of the T-cell acute lymphocytic leukemia protein 1 (TAL1)”



Identification of the binding sites of the T-cell acute lymphocytic leukemia protein 1 (TAL1)

Authors: Mallory Freeberg, Mo Heydarian, Vivek Bhardwaj, Joachim Wolff, Anika Erxleben

Overview

Questions:

- How is raw ChIP-seq data processed and analyzed?
- What are the binding sites of TAL1?
- Which genes are regulated by TAL1?

Objectives:

- Inspect read quality with FastQC
- Perform read trimming with Trimmomatic
- Align trimmed reads with BWA
- Assess quality and reproducibility of experiments
- Identify TAL1 binding sites with MACS2
- Determine unique/common TAL1 binding sites from G1E and Megakaryocytes
- Identify unique/common TAL1 peaks occupying gene promoters
- Visually inspect TAL1 peaks with Trackster

Requirements:

- Introduction to Galaxy Analyses
- Sequence analysis
  - Quality Control: slides - hands-on
  - Mapping: slides - hands-on
- Trackster

Time estimation: 3 hours

Supporting Materials:

Datasets Workflows FAQs Available on these Galaxies

Agenda

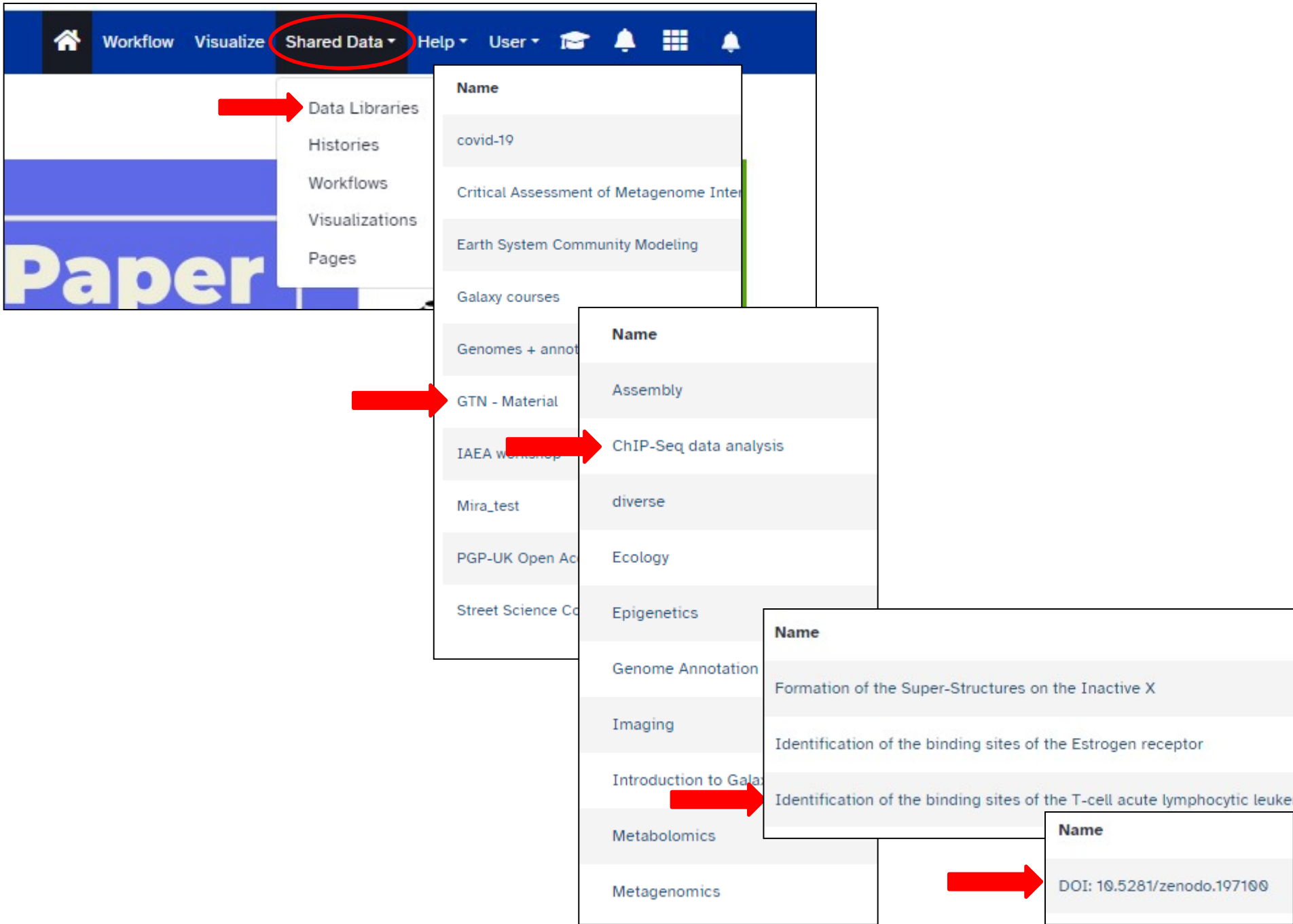
In this tutorial, we will deal with:

1. Quality control
2. Trimming and clipping reads
3. Aligning reads to a reference genome
4. Assessing correlation between samples
5. Assessing IP strength
6. Determining TAL1 binding sites
7. Inspection of peaks and aligned data
    1. Inspection of peaks and aligned data with Trackster
    2. Inspection of peaks and aligned data with IGV
8. Identifying unique and common TAL1 peaks between stages
9. Generating Input normalized coverage files
10. Plot the signal on the peaks between samples
11. Additional optional analyses
    1. Assessing GC bias
12. Conclusion

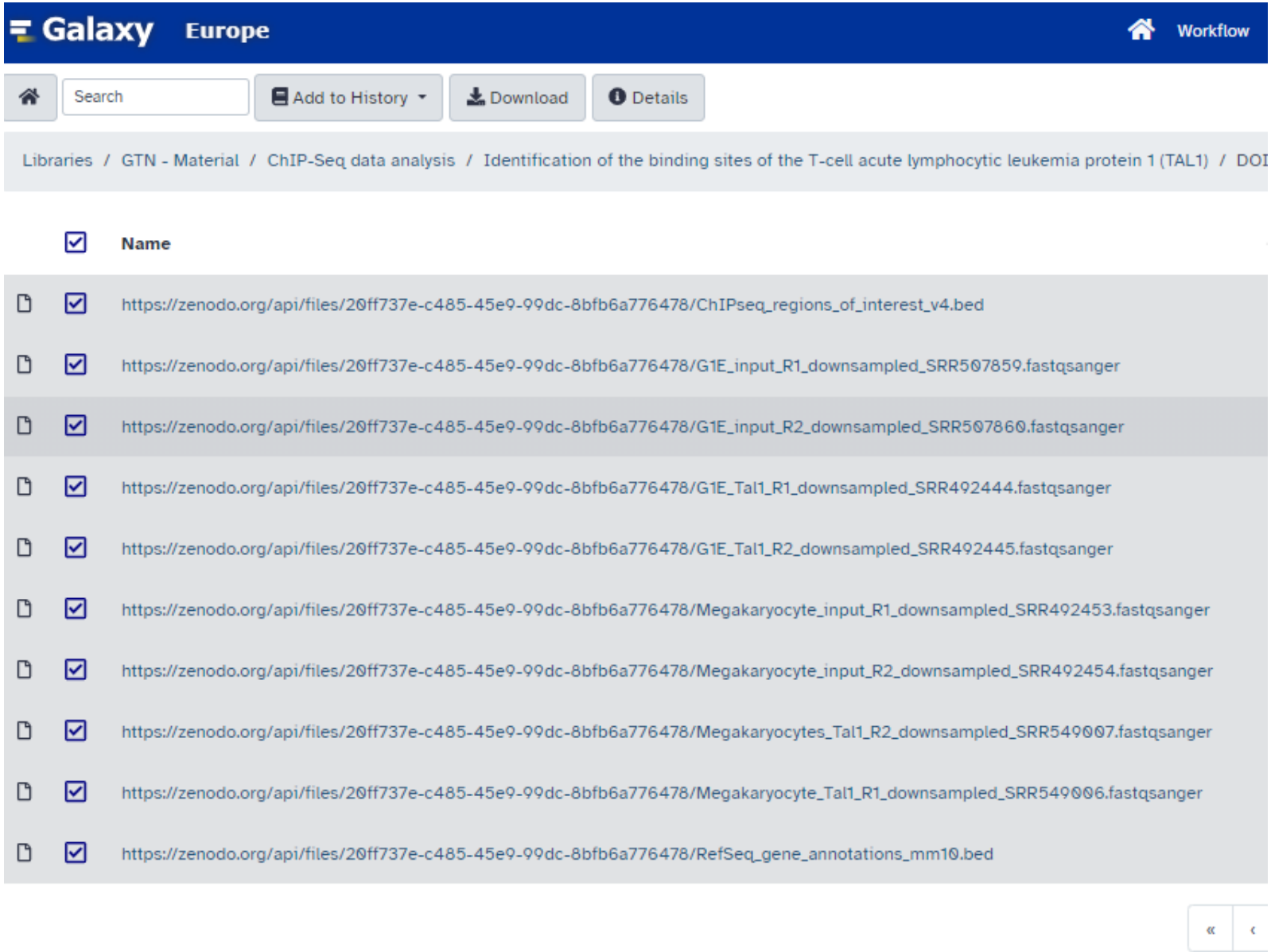
Table 1: Metadata for ChIP-seq experiments in this tutorial. SE: single-end.

Cellular state	Datatype	ChIP Ab	Replicate	SRA Accession	Library type	Read length	Stranded?	Data size (MB)
G1E	ChIP-seq	input	1	SRR507859	SE	36	No	35.8
G1E	ChIP-seq	input	2	SRR507860	SE	55	No	427.1
G1E	ChIP-seq	TAL1	1	SRR492444	SE	36	No	32.3
G1E	ChIP-seq	TAL1	2	SRR492445	SE	41	No	62.7
Megakaryocyte	ChIP-seq	input	1	SRR492453	SE	41	No	57.2
Megakaryocyte	ChIP-seq	input	2	SRR492454	SE	55	No	403.8
Megakaryocyte	ChIP-seq	TAL1	1	SRR549006	SE	55	No	340.3
Megakaryocyte	ChIP-seq	TAL1	2	SRR549007	SE	48	No	356.9

Opção rápida para obter os dados para análise:



Selecionar todos or arquivos e adicionar ao “history” ChIP-seq



https://usegalaxy.eu/training-material/topics/epigenetics/tutorials/tal1-binding-site-identification/workflows/

Aligning reads to a reference genome

To determine where DNA fragments originated from in the genome, the sequenced reads must be aligned to a reference genome. This is equivalent to solving a jigsaw puzzle, but unfortunately, not all pieces are unique. In principle, you could do a BLAST analysis to figure out where the sequenced pieces fit best in the known genome. Aligning millions of short sequences this way, however, can take a couple of weeks. Nowadays, there are many read alignment programs for sequenced DNA, BWA being one of them. You can read more about the BWA algorithm and tool here.

Hands-on: Aligning reads to a reference genome

1. **BWA (Galaxy version 0.7.17.4)**: Run BWA to map the trimmed/clipped reads to the mouse genome.
    - *"Will you select a reference genome..."*: `Use a built-in genome index`
    - *"Using reference genome"*: `Mouse (mus musculus) mm10`
    - *"Select input type"*: `Single fastq`
    - *"Select fastq dataset"*: Select all of the trimmed FASTQ files

2. Rename files to reflect the origin and contents

Determining TAL1 binding sites

Now that BWA has aligned the reads to the genome, we will use the tool MACS2 to identify regions of TAL1 occupancy, which are called "peaks". Peaks are determined from pileups of sequenced reads across the genome that correspond to where TAL1 binds.

MACS2 will perform two tasks:

1. Identify regions of TAL1 occupancy (peaks).
2. Generate bedGraph files for visual inspection of the data on a genome browser.

More information about MACS2 can be found in Zhang *et al.* 2008.

Hands-on: Determining TAL1 binding sites

1. **MACS2 callpeak** (Galaxy version 2.1.1.20160309.6): Run MACS2 callpeak with the aligned read files from the previous step as Treatment (TAL1) and Control (input).
    - *"Are you pooling Treatment Files?"*: `Yes`
    - *"ChIP-Seq Treatment File"*: Select all of the replicate ChIP-Seq treatment aligned BAM files for one cell type
    - *"Do you have a Control File?"*: `Yes`
    - *"Are you pooling Control Files?"*: `Yes`
    - *"ChIP-Seq Control File"*: Select replicate ChIP-Seq control aligned BAM files for the same cell type
    - *"Format of Input Files"*: `Single-end BAM`
    - *"Effective genome size"*: `M. musculus`
    - *"Additional Outputs"*: Select `Peaks as tabular file (compatible wih MultiQC)`, `Peak summits`, `Scores in bedGraph files (--bdg)`
2. Rename files to reflect the origin and contents.
3. Repeat for the other cell type.

Identifying unique and common TAL1 peaks between stages

We have processed ChIP-seq data from two stages of hematopoiesis and have lists of TAL1-occupied sites (peaks) in both cellular states. The next analysis step is to identify TAL1 peaks that are *shared* between the two cellular states and peaks that are *specific* to either cellular state.

Hands-on: Identifying unique and common TAL1 peaks between states

1. **bedtools Intersect intervals (Galaxy version 2.29.0)**: Run bedtools Intersect intervals to find peaks that exist both in G1E and megakaryocytes.
    - *"File A to intersect with B"*: Select the TAL1 G1E narrow peaks BED file
    - *"File B to intersect with A"*: Select the TAL1 Megakaryocytes narrow peaks BED file
    - Running this tool with the default settings will return overlapping peaks of both files.

2. **bedtools Intersect intervals (Galaxy version 2.29.0)**: Run bedtools Intersect intervals to find peaks that exist only in G1E.
    - *"File A to intersect with B"*: Select the TAL1 G1E narrow peaks BED file
    - *"File B to intersect with A"*: Select the TAL1 Megakaryocytes narrow peaks BED file
    - *"Report only those alignments that **do not** overlap the BED file"*: **Yes**

3. **bedtools Intersect intervals (Galaxy version 2.29.0)**: Run bedtools Intersect intervals to find peaks that exist only in megakaryocytes.
    - *"File A to intersect with B"*: Select the TAL1 Megakaryocytes narrow peaks BED file
    - *"File B to intersect with A"*: Select the TAL1 G1E narrow peaks BED file
    - *"Report only those alignments that **do not** overlap the BED file"*: **Yes**

4. Rename files to reflect the origin and contents.

Instruções para o relatório

1) Descreva sucintamente os passos usados na análise de ChIP-seq

.

2) Responda as seguintes perguntas, explicando como chegou a essas conclusões:

Quantos picos TAL1 são comuns às células G1E indiferenciadas e aos megacariócitos?

Quantos picos são exclusivos das células G1E?

Quantos picos são exclusivos dos megacariócitos?