Regulatory Genomics and Epigenomics

Izaskun Mallona and Tuncay Baubec

17th October 2023

Swiss Institute of Bioinformatics

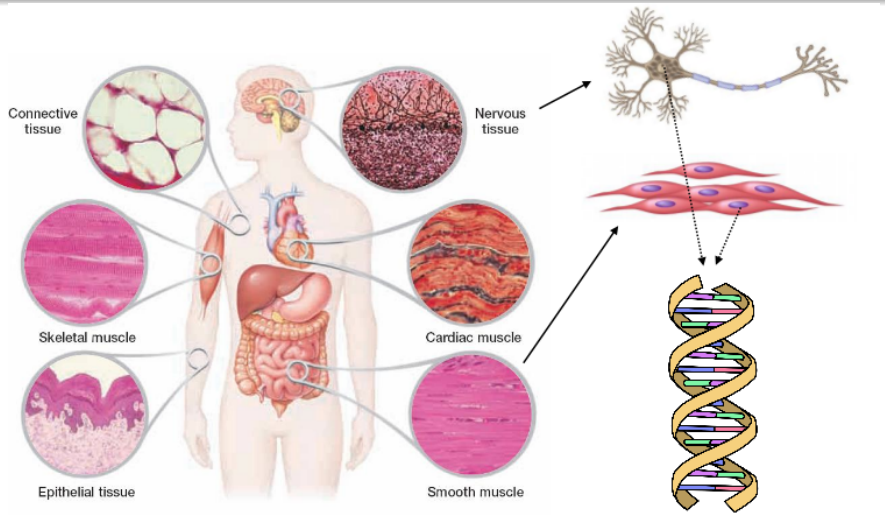
Structure

- Lecture:
  - Part I: Introduction to the field and experimental procedures
  - Part II: Computational challenges and strategies
- Example data to browse (by Tuncay Baubec)
- Questions: izaskun.mallona@mls.uzh.ch

Objectives: questions to address

- What is gene regulation? is the same as genome regulation?
- Why does it matter?
- How can we analyze it?
  - Which kind of data do we get experimentally?
  - How do process these data (computational biology)?

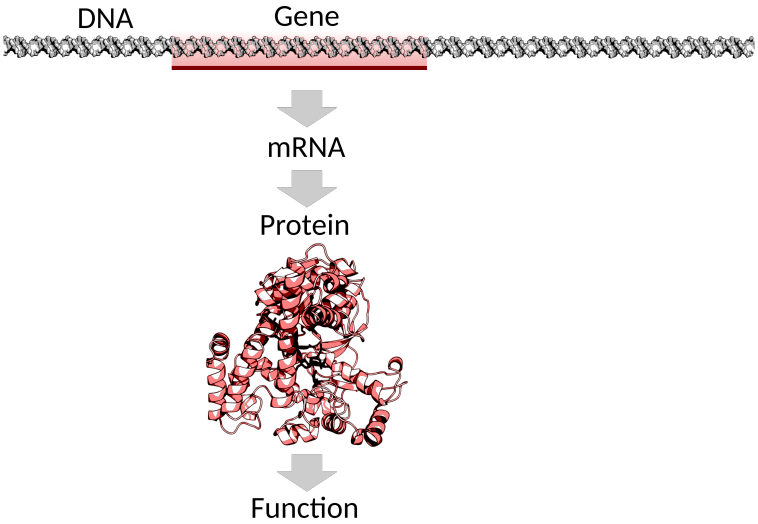
Regulatory diversity: same DNA, different phenotypes



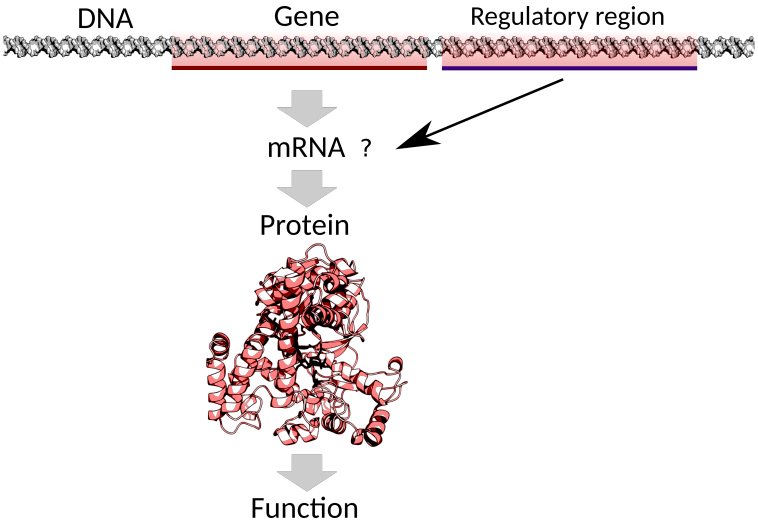
~200 cell types with different properties

same genetic material

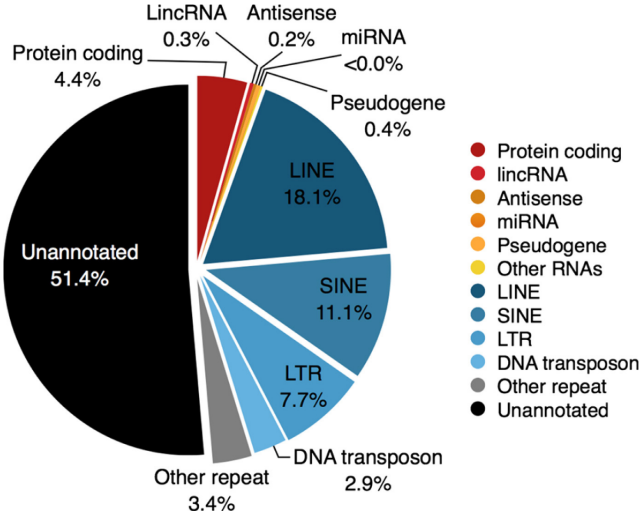
Genome, transcriptome and proteome: a simplified model



Regulome, genome, transcriptome and proteome

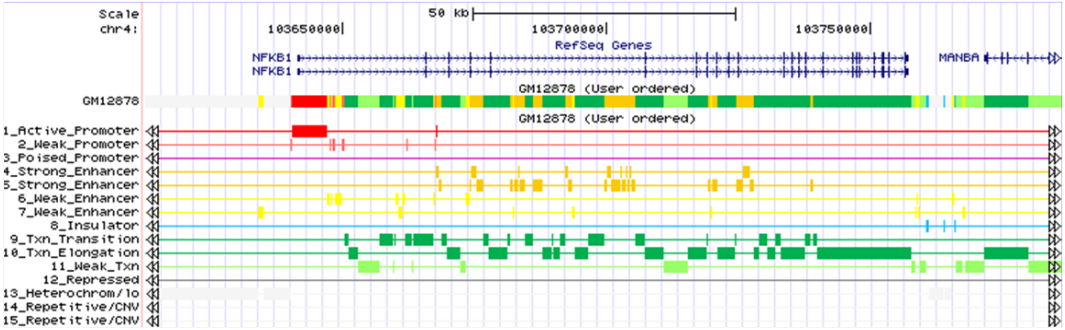


What's the genome content? Are coding regions abundant?



Hutchins and Pei 2015

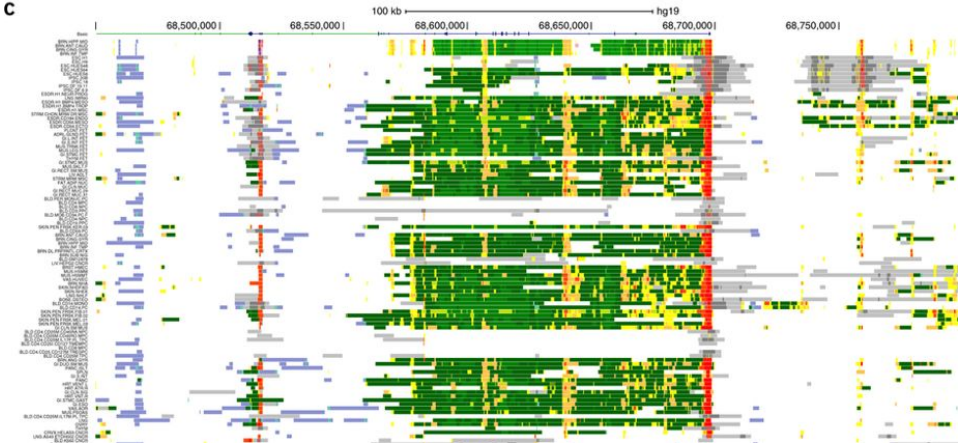
Genomic information encoding: ENCODE



Ernst and Kellis 2012 ChromHMM

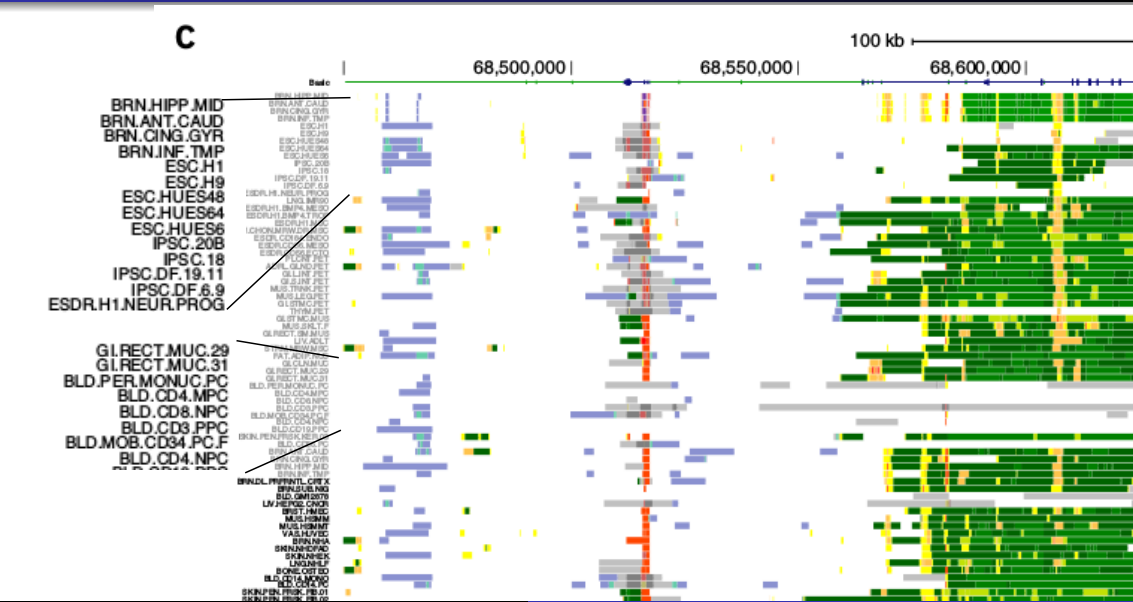
(we will come back to this slide at the end of the talk)

The regulatory genome (segmentation by ChromHMM)



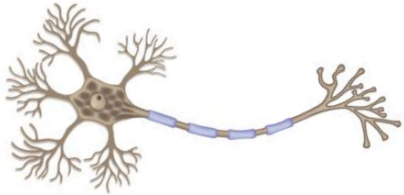
- 200 kbp human genome coordinates (X axis) chromatin states for different cell types (Y axis). Green indicates trancription (Ernst 2017)

Zooming in the regulatory genome (segmentation by ChromHMM)

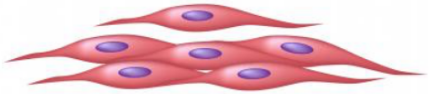


Gene regulation

- 20,000-25,000 genes (exons are only 1% of the genome)
- some genes are required in all cell types, but some are relevant only for specific cells



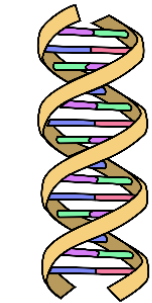
Gene A = ON
Gene B = OFF



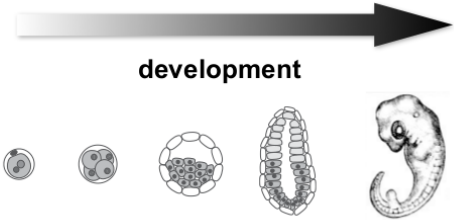
Gene A = OFF
Gene B = ON

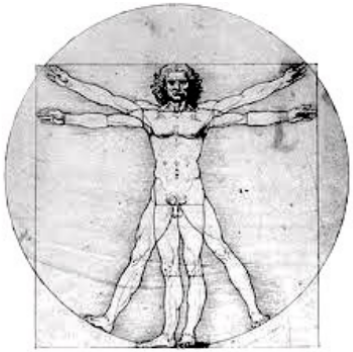
Genome regulation

- Genome wide language for activation (switching on) and repression (switching off) genes
- Flexibility: spatial and temporal regulation



DNA

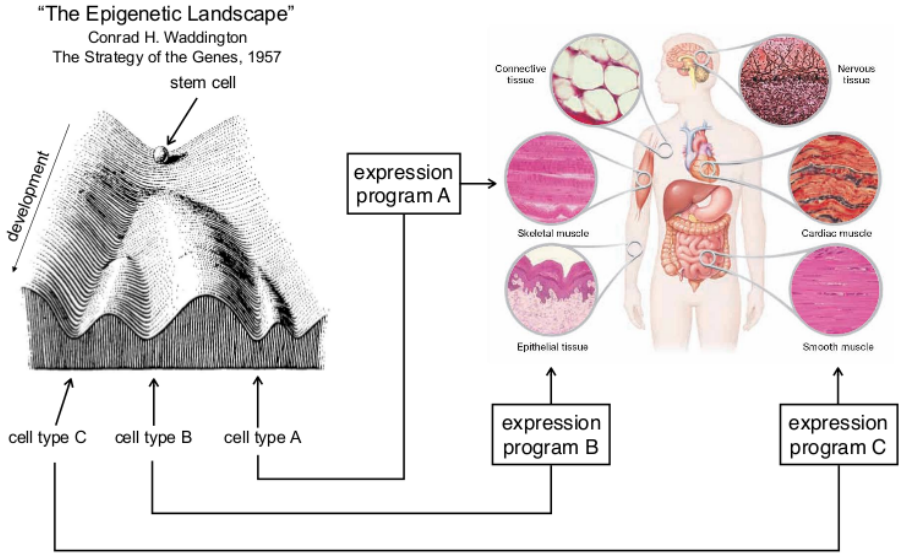




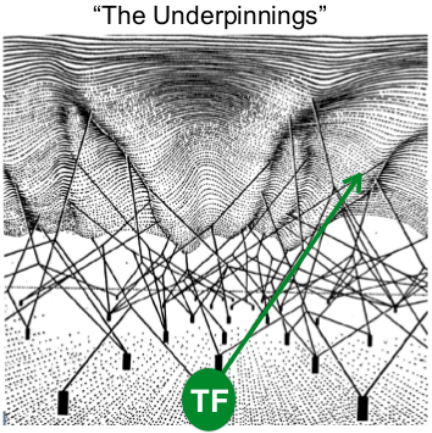
One genome

multiple cell types

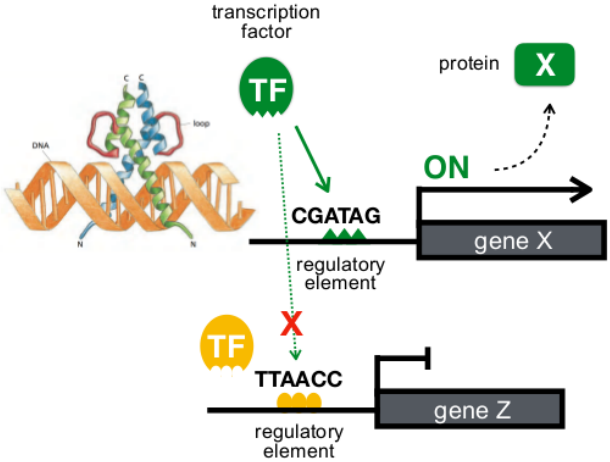
The Epigenetic Landscape (Waddington)



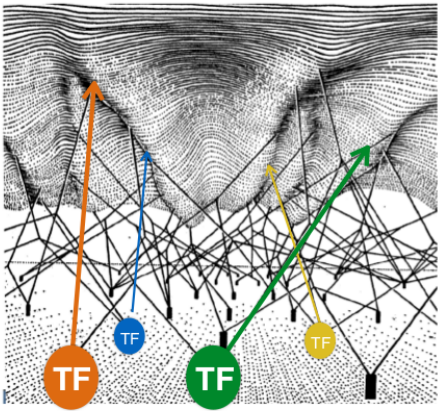
Determining gene expression programs: transcription factors



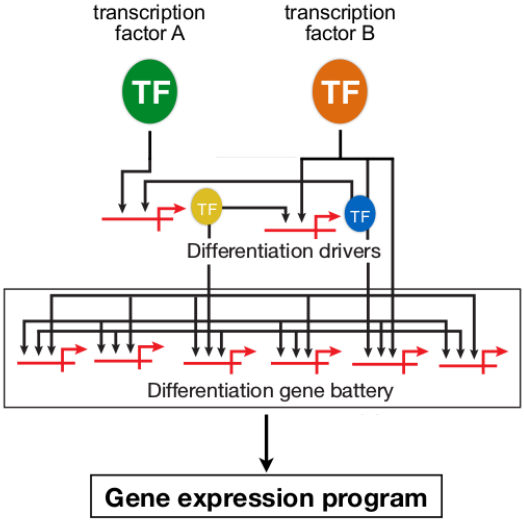
Conrad H. Waddington
The Strategy of the Genes, 1957



Transcription factors



Conrad H. Waddington
The Strategy of the Genes, 1957



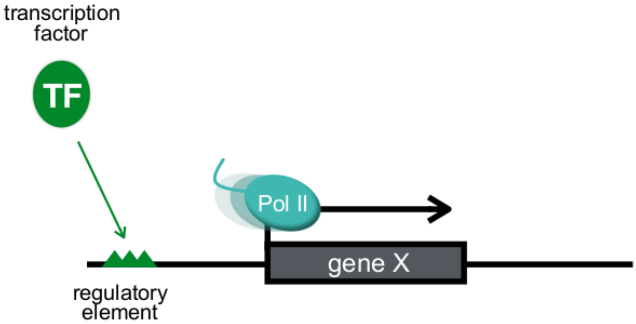
Adopted from Davidson Nature 2010

Transcription factor binding sites

- Favourable binding sites are specific for each transcription factor
- Consensus represented as sequence logos
- Sequence logos stack letters (bases) whose relative sizes indicate their frequency in the sequences

Gene	Motif	q-value (Benjamini)
Oct4		0.0071
Klf-4		0.0232
Nanog		0.0573
Sox2		0.0799

Transcription factors switches



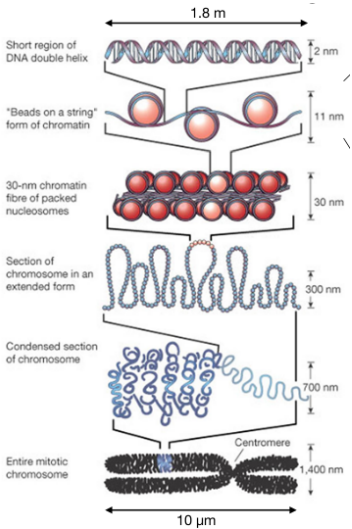
- Gene (defined **start** and end)
- Transcription machinery (RNA Polymerase + complex)
- Transcription factors (repressor/activator)
- Regulatory DNA sequences

= switch

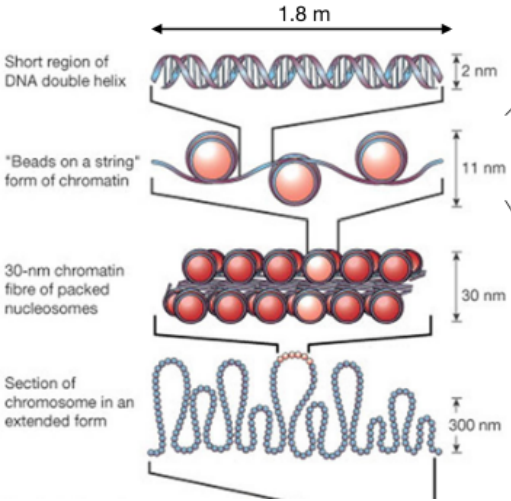
- **Chromatin & epigenetic modifications**

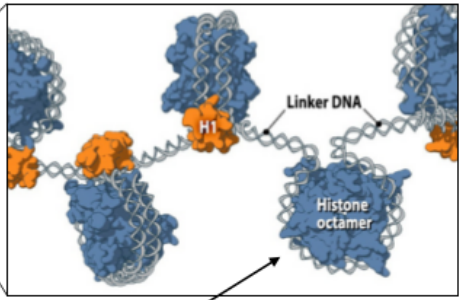
DNA folding levels

- Genome has a 3D organization inside the nucleus: folding 1.8 m of DNA vs nucleus diameter of 4-6 microns
- DNA folding (chromatin accessibility) influences binding of transcription factors to DNA
- With folding, what makes an enhacer close to a gene is not (only) being next in DNA sequence



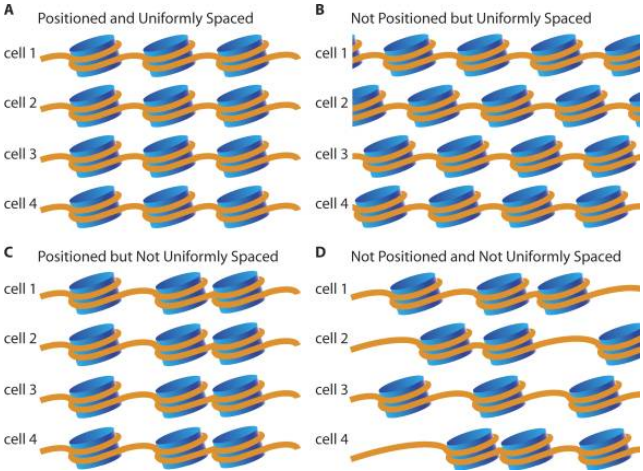
Nucleosomes 3D





Nucleosome: basic subunit of chromatin
- contains DNA and histones (proteins)

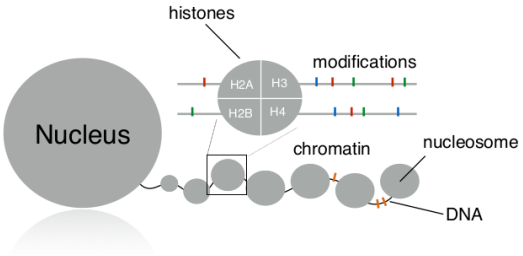
Nucleosomes positioning



Valouev, 2008

Chromatin modifications

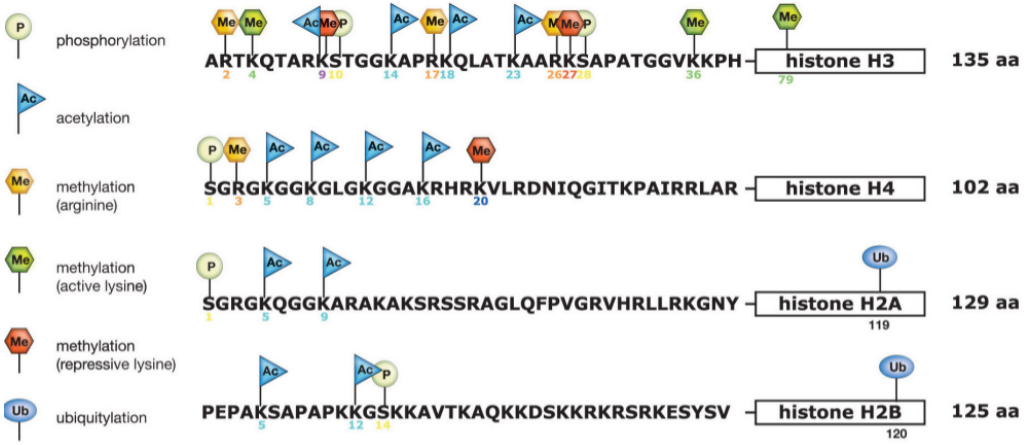
- Chromatin compaction varies between cells
- Chromatin suffers modifications in both histones (proteins) and DNA itself



on histones:
- majority on histone H3 and H4 N-terminal tail
- depending on the type of modification and position: activating or repressing

on DNA:
- methylation (and oxidative derivates)
- mainly on cytosines, but also on adenines

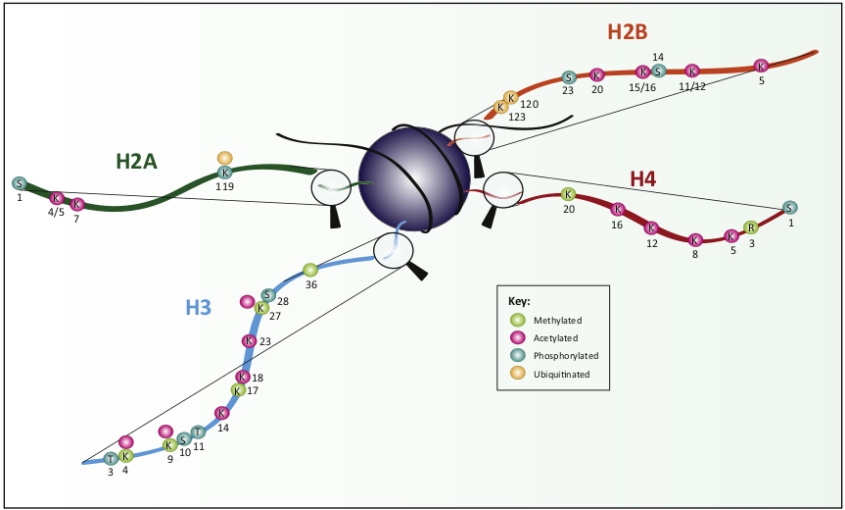
Chromatin compaction and modifications: histones



Chromatin compaction and modifications: histones

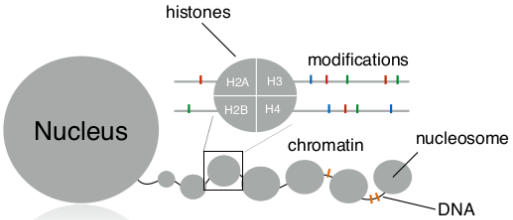
- Naming: Histone name - aminoacid number - decoration
- e.g. H3K4me3: histone 3, lysine 4, trimethyl (has 3 methyl groups)

Chromatin compaction and modifications: histone tails



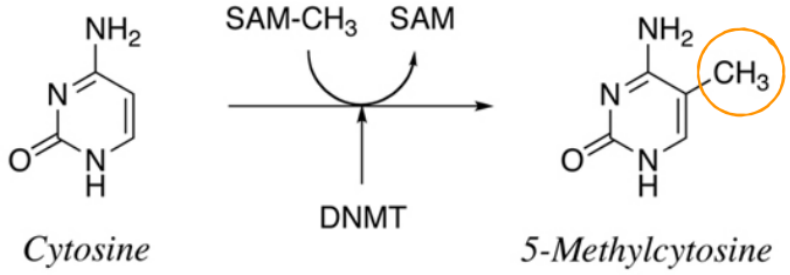
(Lawrence, Daujat, Schneider 2015)

Chromatin compaction and modifications: DNA



on histones:
- majority on histone H3 and H4 N-terminal tail
- depending on the type of modification and AA residue: activating or repressing

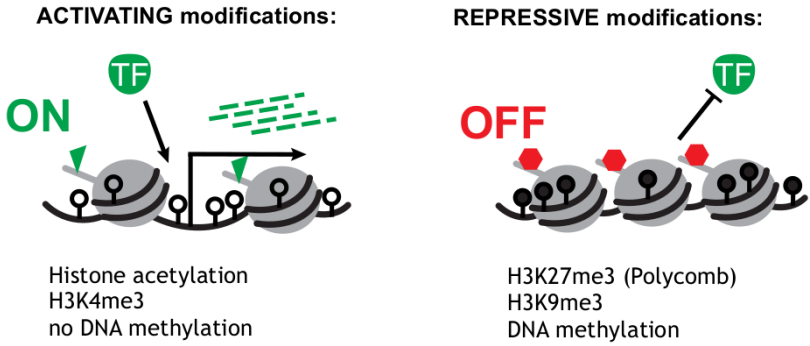
on DNA:
- methylation of cytosines
- repressive mark

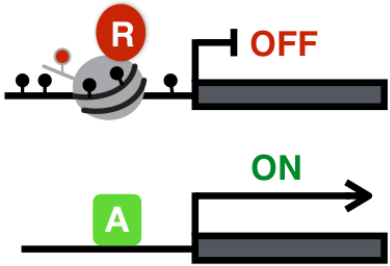


Cytosine *5-Methylcytosine*

Chromatin compaction impact in regulation

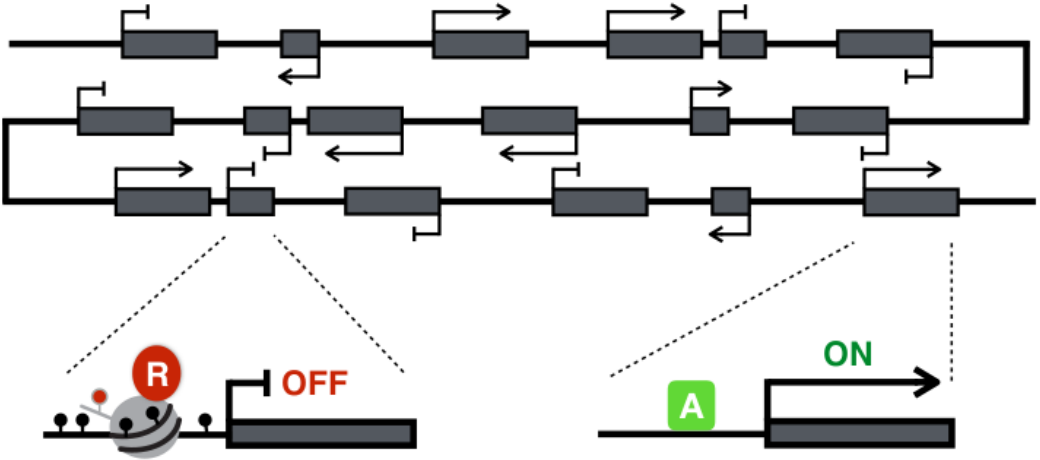
- Chromatin packaging and modifications influence the accessibility and transcriptional output of the genome
- How?





- which genes are active in cell type X ?
- which factors regulate their activity ?
- what are their chromatin states ?
- which regulatory sequences are present ?

Genome-wide localisation analysis



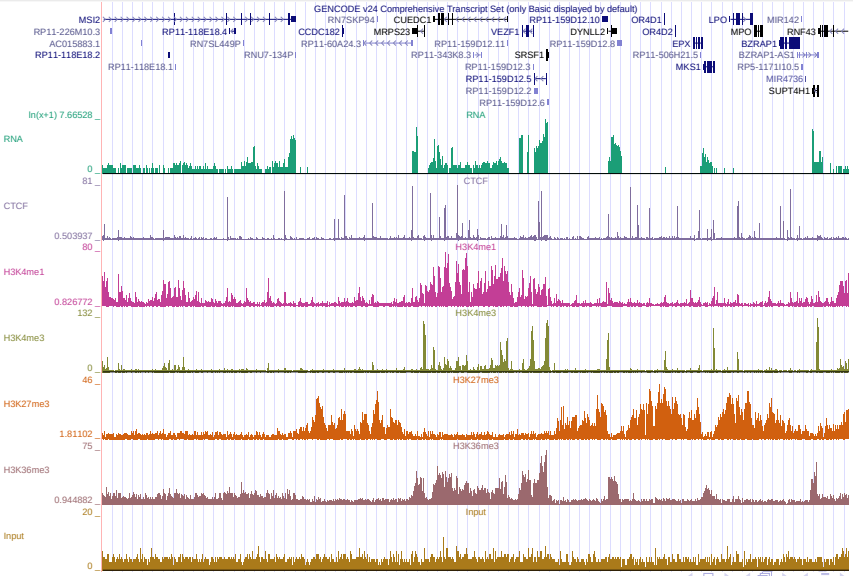
Localising what?

- Transcription?
- Transcription factor binding?
- Specific sequences recognized by transcription factors (i.e. transcription factor binding motifs)?
- Histone modifications?
- DNA modifications?

Informatics: genome-wide representations

- We need data representations to integrate and visualize these regulatory layers on top the entire genome
- (as well as methods to detect the molecular fingerprints)

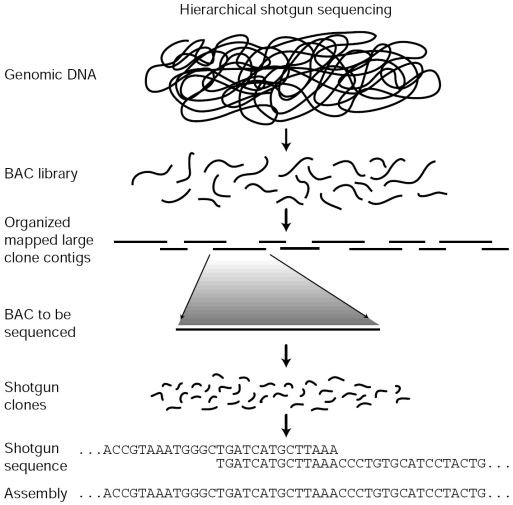
Meeting data and biology: coordinate-based output



How to store these data?

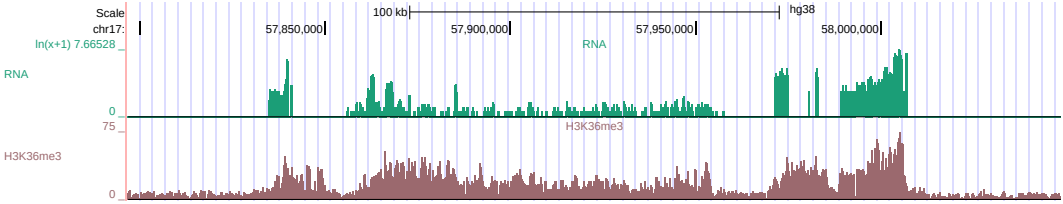
- Quantitative, unbiased readouts for specific genomic coordinates
- Data storage costs: each human genome is 3 billion basepairs
- How to store the active and inactive marks of each cell type?
- Does it make sense to have multiple copies of the same DNA tagged with different labels?
- Data standards get rid of sequences and, rather, use coordinates of reference genomes

Human reference genome (Nature 2001)



Genome coordinates: lingua franca of genomic annotations

- Components needed to stack information layers on top of a genome:
  - Name of the assembly (hg38)
  - Chromosome (or scaffold)
  - Start, ends, scores and the name of whatever we are measuring (i.e. RNA levels and H3K36me3, an histone modification linked to transcription).



Keep it simple: storing genomic coordinates

- BED (Browser Extensible Data) files define genomic loci as plain text files so they don't store sequences but, rather, were the features are
- BED3: 3 tab separated columns, chromosome (scaffold), start, end
- BED6: BED3 plus name, score, strand

BED3: simplest coordinate-based file format

- How to store two features of 1 kbp each (could be active regions in a given cell type) located at chromosome 22 and starting at nt number 1000 and 8000, respectively?

```
chr22 1000 2000
chr22 8000 9000
```

BED6

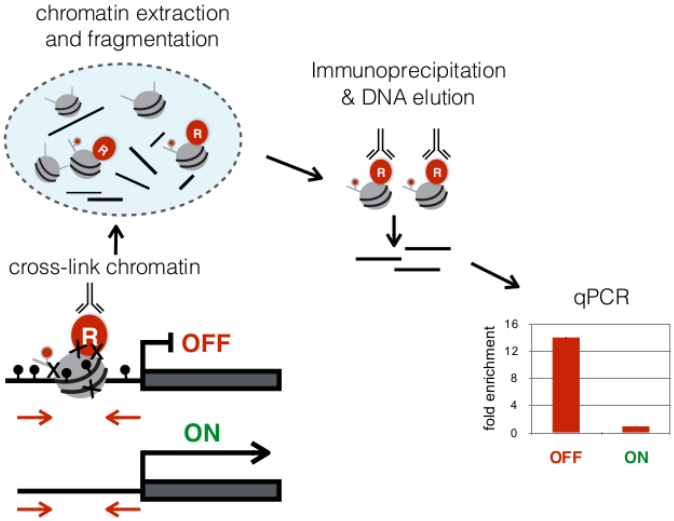
- How to store the strand, name of the feature and a score as well?
- BED6 format: 6 tab separated columns, chromosome (scaffold), start, end, name, score, strand

```
chr22 900  1100 promoter  1000 +
chr22 1000 1200 enhancer  1000 +
chr22 1100 6000 gene_body 1000 +
```

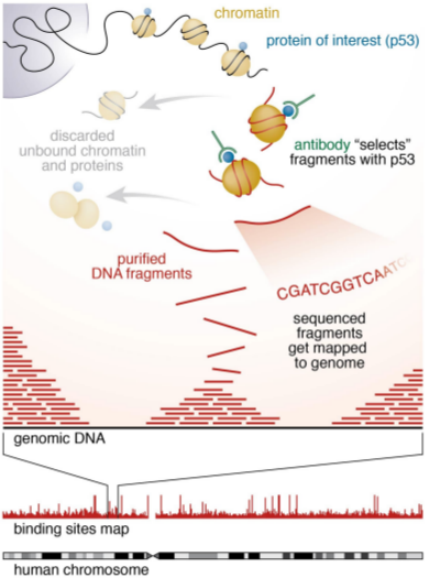
Molecular methods: ChIP

- Still, were do data come from?
- ChIP (chromatin immunoprecipitation) analyze protein interactions with DNA
- ChIP-seq combines ChIP with parallel DNA sequencing to identify the binding sites of DNA-associated proteins

Localisation analysis by ChIP

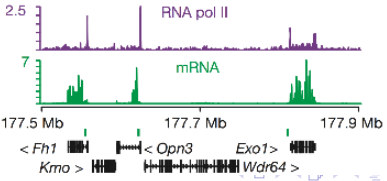


Localisation analysis by ChIP-Seq



Chromatin Immunoprecipitation - sequencing
method to identify genomic location of protein of interest (e.g. TF, RNA Pol2) or histone modifications

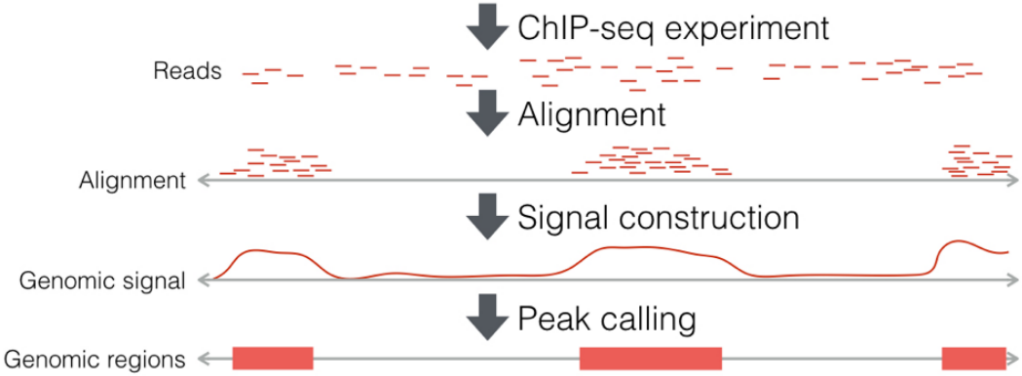
1. proteins are fixed to chromatin by formaldehyde (crosslinking)
2. chromatin is sheared to 100-300bp (ultrasound or enzymes)
3. specific antibodies enrich pieces of DNA bound by protein of interest
4. enriched DNA is purified and sequenced
5. usually 20-100 mio sequences are obtained from one experiment = "reads"
6. sequences reads are "mapped" back to the genome to identify their position along the chromosome
7. signal intensity indicates localisation frequency



ChIP-Seq data flow

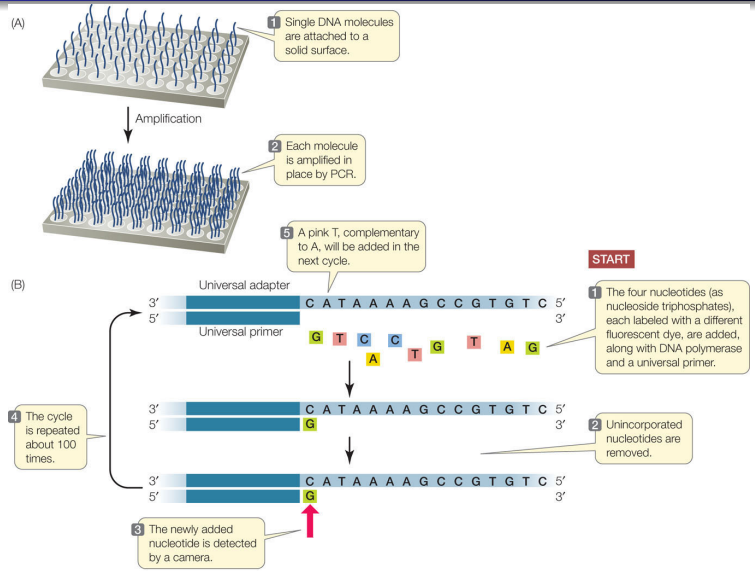
- High-throughput short read sequencing
- Read quality control
- Mapping the sequences back to the reference genome
- Mapping quality control
- Summarization into coordinate-based files

ChIP-Seq data flow



regulatory-genomics.org

High-throughput short read sequencing

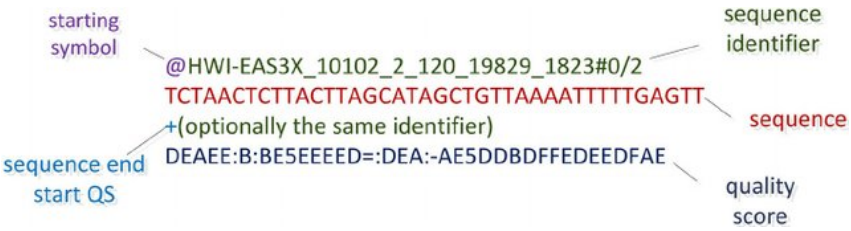


Macmillan learning

Read quality control

- Reads have a Phred quality score (of a given base Q) as defined by $Q = -10 \log_{10} P$
- Data stored as FASTQ files (each with dozens/hundreds million sequences)

FASTQ records

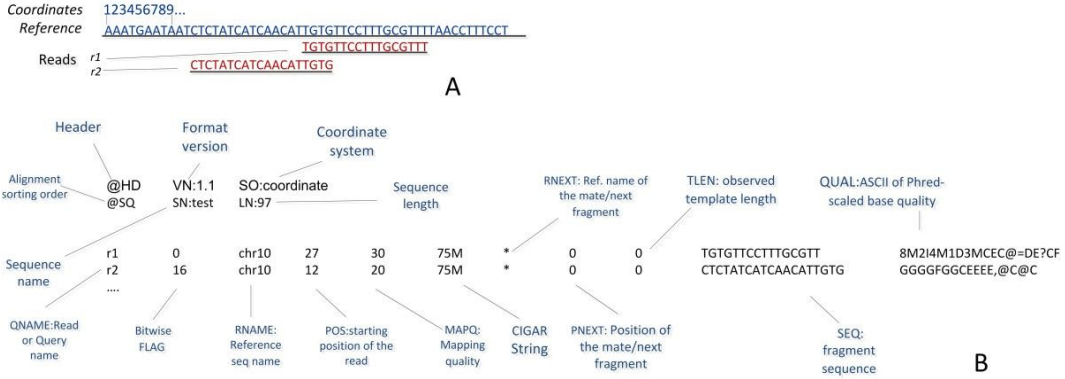


Pavlopoulos et al 2013

Mapping the sequences to a reference

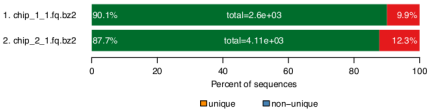
- Assigning each read a location of a reference genome
  - What if the sequences are not identical?
  - How reliable is the reference genome?

Mapping the sequences to the reference genome: SAM files



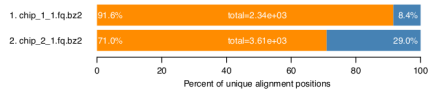
Pavlopoulos et al 2013

Mapping challenges: mappability and repeats



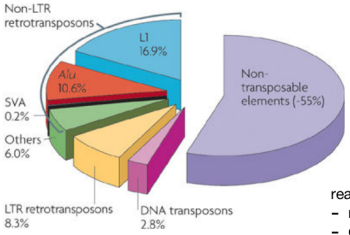
Mapability
Indicates percentage of mapped reads to the genome per sample.

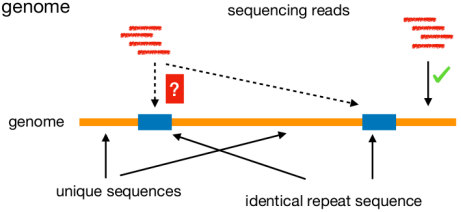
mapped unmapped



Uniqueness
Indicates percentage of uniquely mapped reads to the genome per sample.

Human genome: > 40 % of the mammalian genome contains repetitive sequences





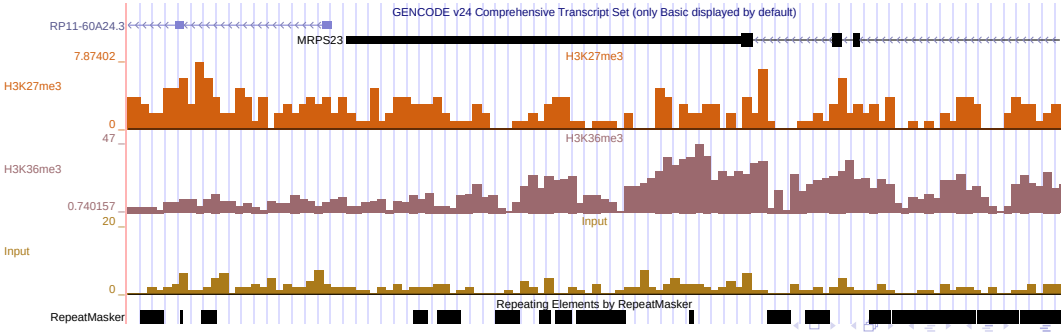
reads mapping to multiple sites:
- randomly assigned to a site & weighted by the number of positive hits
- discarded, depending on the cutoff of multiple mapping possitions

Mapping challenges: SNPs

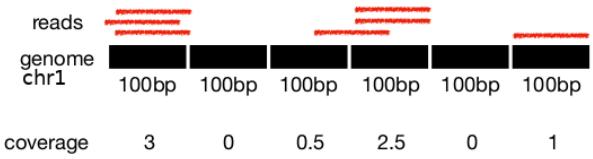
- Genetic variation brings mismatches into play

Summarization into coordinate-based files

- Assigning a value (score) to each readout, i.e. gene expression, binding of a transcription factor, amount of DNA methylation etc to specific genomic coordinates
- BED files
- And/or other optimized data storage options that provide a value along the genome using a fixed interval: Wiggle files



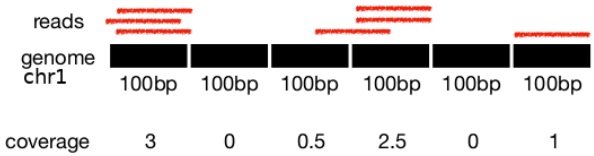
Wig vs BED files: continuous data



- BED format (note that BED-like format is flexible to specify any interval length, not only 100 bp)

```
chr1    1     100   3
chr1    101   200   0
chr1    201   300   0.5
chr1    301   400   2.5
chr1    401   500   0
chr1    501   600   1
```

Wig vs BED files: continuous data



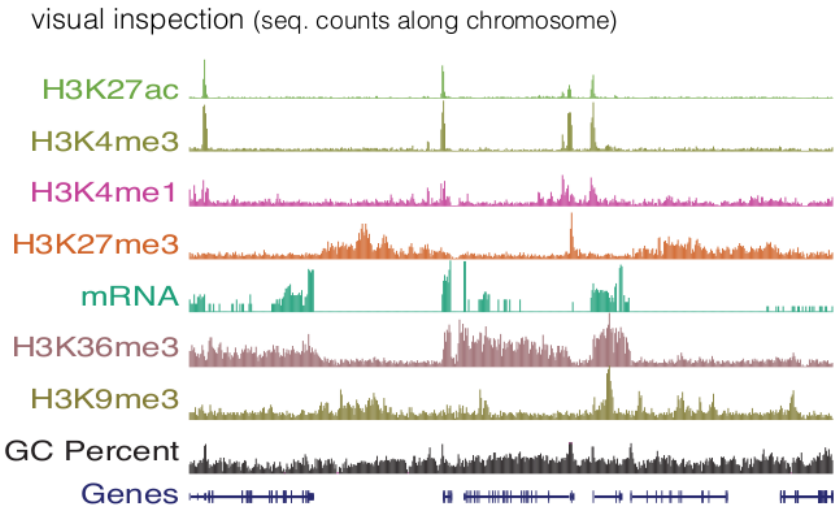
- Wiggle format: getting rid of redundant information

```
variableStep chrom=chr1 span=100
3
0
0.5
2.5
0
1
```

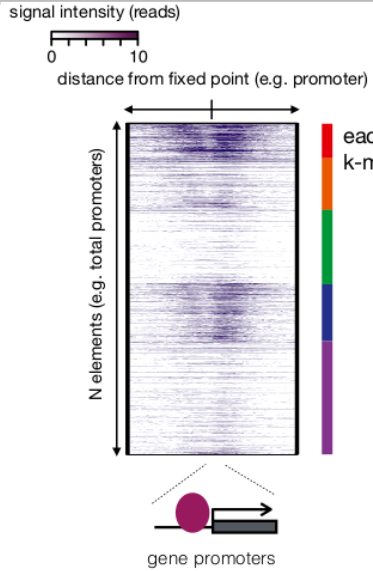
Data visualization and interpretation

- Once the data is mapped and stored in standard file formats analysis follows, including:
  - Visual inspection
  - Regions of interest and regions clustering
  - Peak calling
  - Extracting sequence information (sequence logos)
  - Machine learning

Visual inspection: peak association

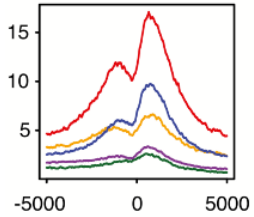


How to interpret these data? Regions of interest

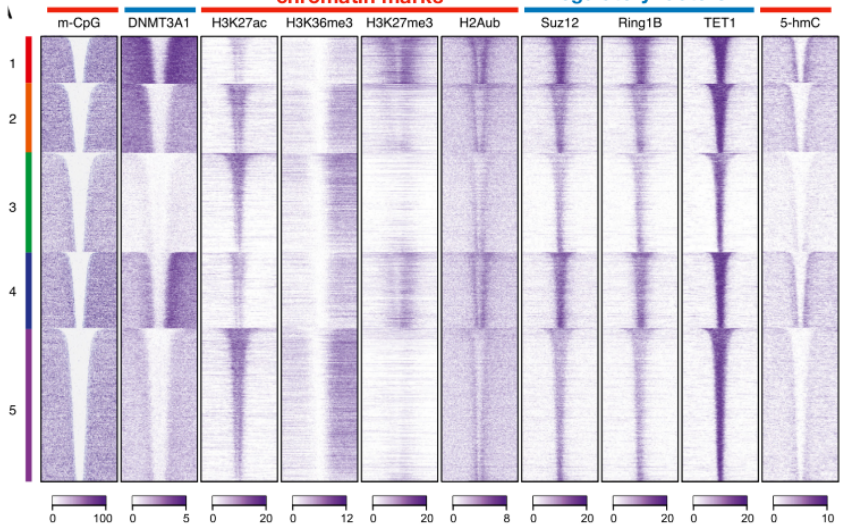


each row = one gene promoter
k-means clustering based on signal distribution

plot average signal of each cluster



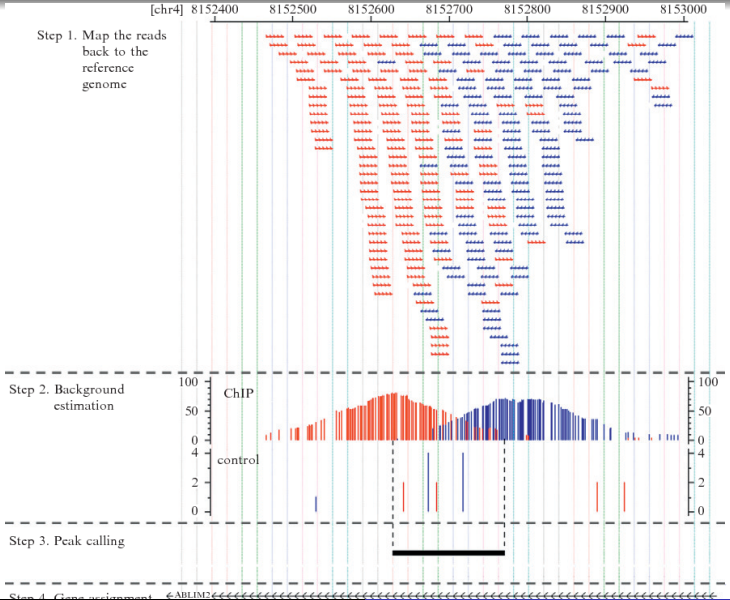
How to interpret these data? Regions of interest

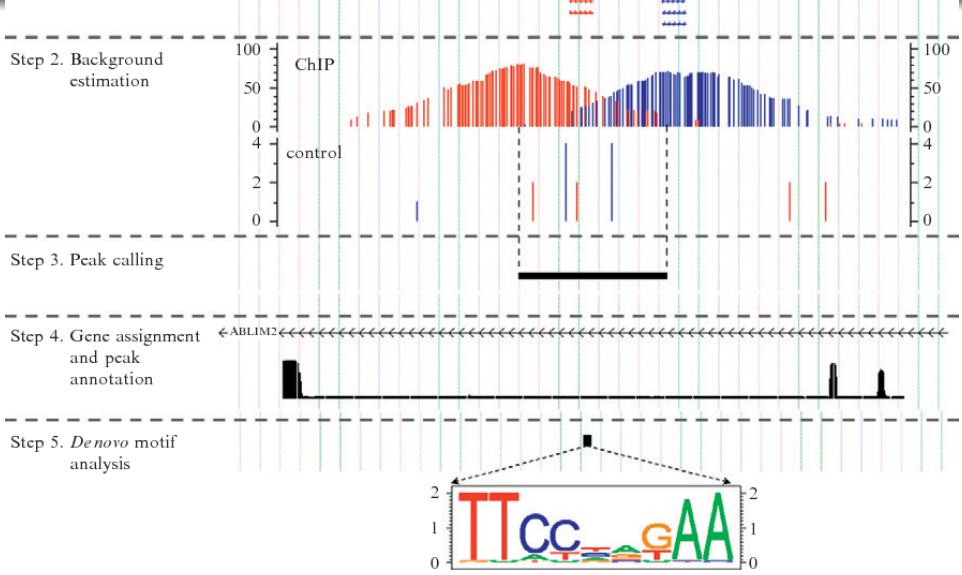


Extracting patterns: getting transcription factor binding sites

- Once the ChiP-Seq data is mapped and the coordinates called, we can extract the sequence patterns that sustain the transcription factor binding

Motif analysis after peak calling (Ma et al, 2011)

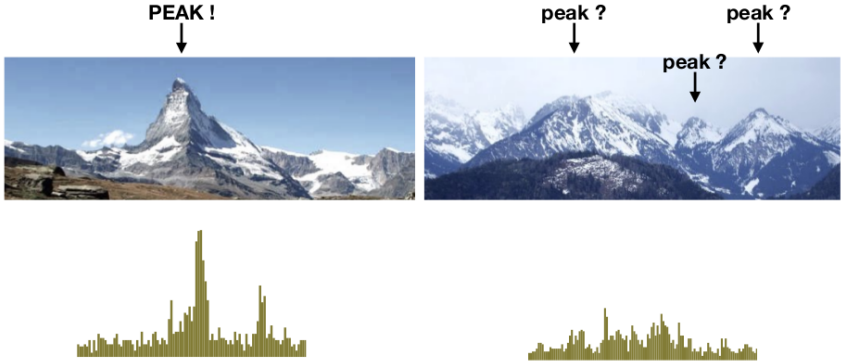



Step 2. Background estimation
ChIP
control
100
50
0
4
2
0
Step 3. Peak calling
Step 4. Gene assignment and peak annotation
ABLIM2
Step 5. *Denovo* motif analysis
2
1
0

Ma et al, 2011

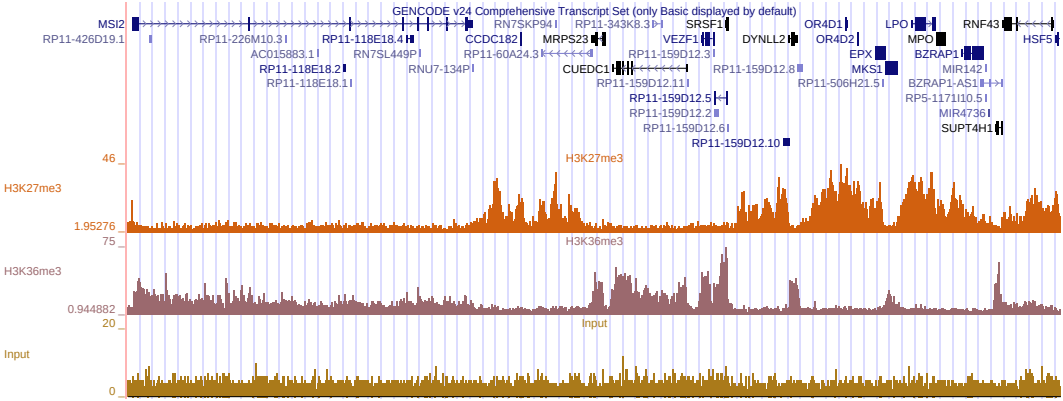
Is peak calling that simple?

- How to get the exact locus to which a transcription factor binds?
- Peak calling: detecting peaks while accounting for noise



Peak calling: input samples

- Input/mock data: just chromatin without immunoprecipitation, or ChiP of IgG (i.e. no treatment)
- Account for noise during the mapping/peak calling process



Machine learning

- Can we learn chromatin states and correlate with regulatory functions using machine learning methods?
- Which data can we feed into them?

Machine learning: genome segmentation using ChromHMM

ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

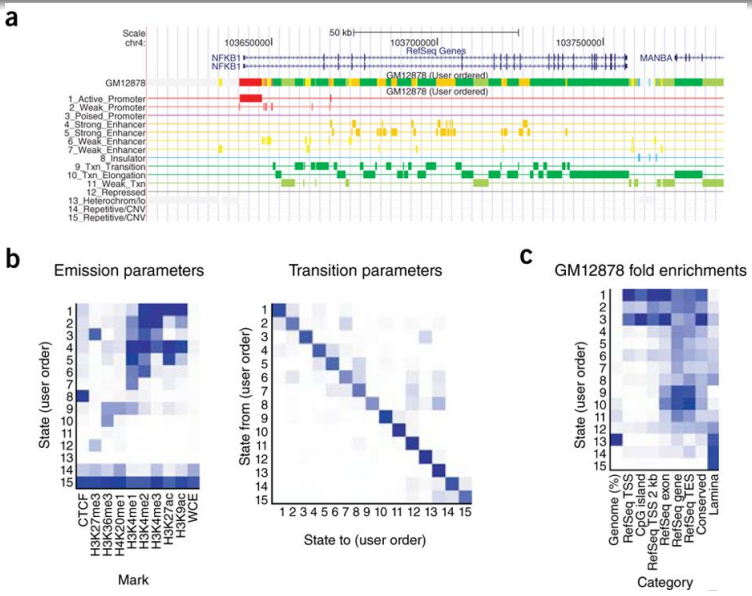
Nature Methods **9**, 215–216 (2012) | Download Citation

To the Editor:

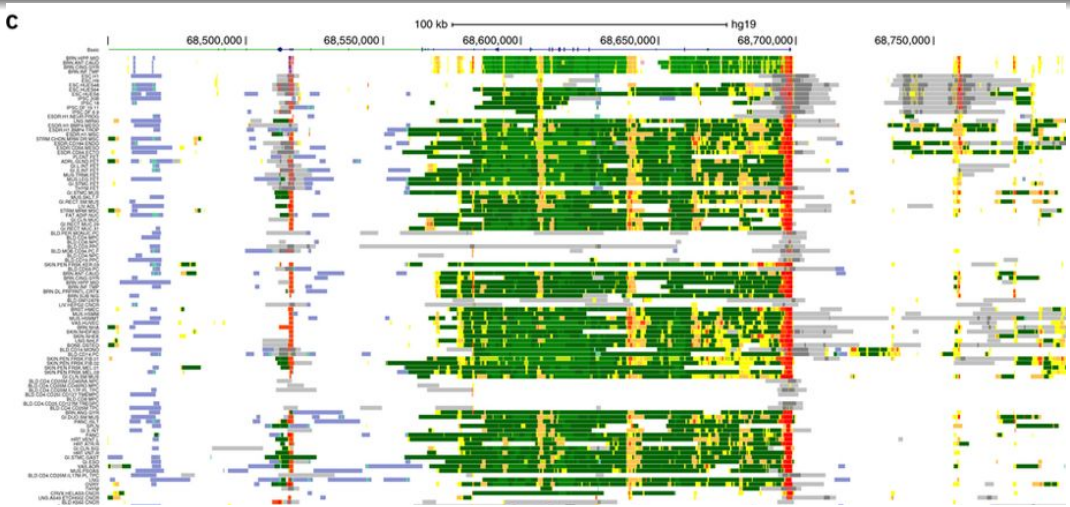
Chromatin-state annotation using combinations of chromatin modification patterns has emerged as a powerful approach for discovering regulatory regions and their cell type–specific activity patterns and for interpreting disease-association studies[1,2,3,4,5]. However, the computational challenge of learning chromatin-state models from large numbers of chromatin modification datasets in multiple cell types still requires extensive bioinformatics expertise. To address this challenge, we developed ChromHMM, an automated computational system for learning chromatin states, characterizing their biological functions and correlations with large-scale functional datasets and visualizing the resulting genome-wide maps of chromatin-state annotations.

- ChromHMM: released in 2012
- Still widely used: e.g. latest 2021 ENCODE's release (example paper), and resource

Machine learning: genome segmentation using ChromHMM



Cell types segmentation by ChromHMM



- Rows: cell. Darkgreen, transcription; red, TSS; blue, heterochromatin; yellow-green, enhancer; gray, repressed.

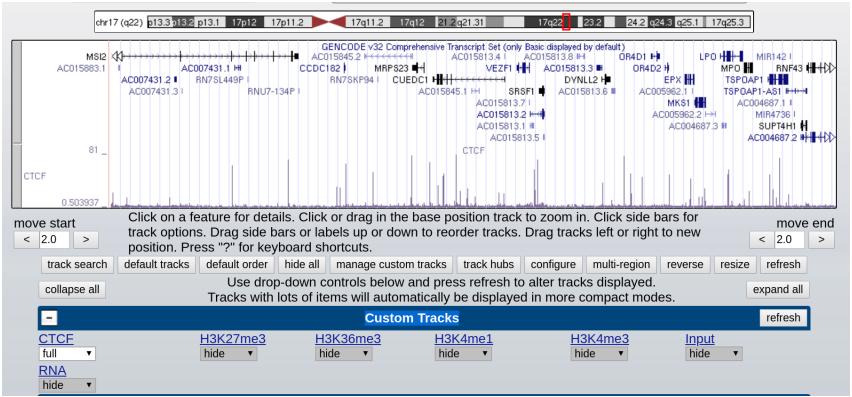
Sum up

- Regulatory genomics and epigenomics study non-coding regulatory regions in the genome
- Regulatory genomics characterizes regulatory diversity (i.e. during development and/or cell diversity)
- Regulatory genomics focus on computational problems including the identification of regions that potentially control the regulation of specific genes, i.e.
  - Identification of regulatory regions
  - Identification of over-represented sequence motifs in sets of regulatory regions
  - Machine learning of complex signatures

UCSC Genome Browser session

- Regulatory data track (hg38 human assembly; by Tuncay Baubec)

http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=helitron&hgS_otherUserSessionName=hg38_BI0390



Extra resources

- Introduction to Regulatory Genomics and Epigenomics https://simons.berkeley.edu/talks/regulatory-genomics-epigenomics : longer/in depth overview of to the field

Extra resources

- Introduction to Regulatory Genomics and Epigenomics https://simons.berkeley.edu/talks/regulatory-genomics-epigenomics : longer/in depth overview of to the field
- Contact izaskun.mallona@mls.uzh.ch