

# Learning Chromatin States from ChIP-seq data

Luca Pinello

GC Yuan Lab



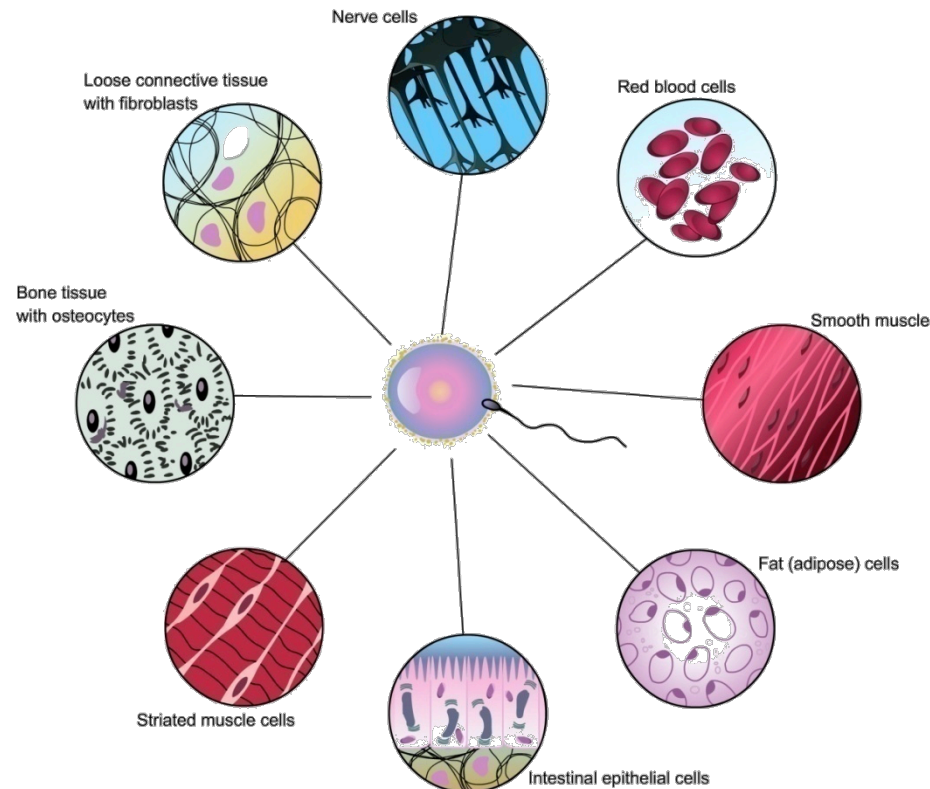
**HARVARD**  
**T.H. CHAN**  
SCHOOL OF PUBLIC HEALTH

# Outline

- Chromatin structure, histone modifications and combinatorial patterns
- How to segment the genome in chromatin states
- How to use ChromHMM step by step
- Further references

# Epigenetics and chromatin structure

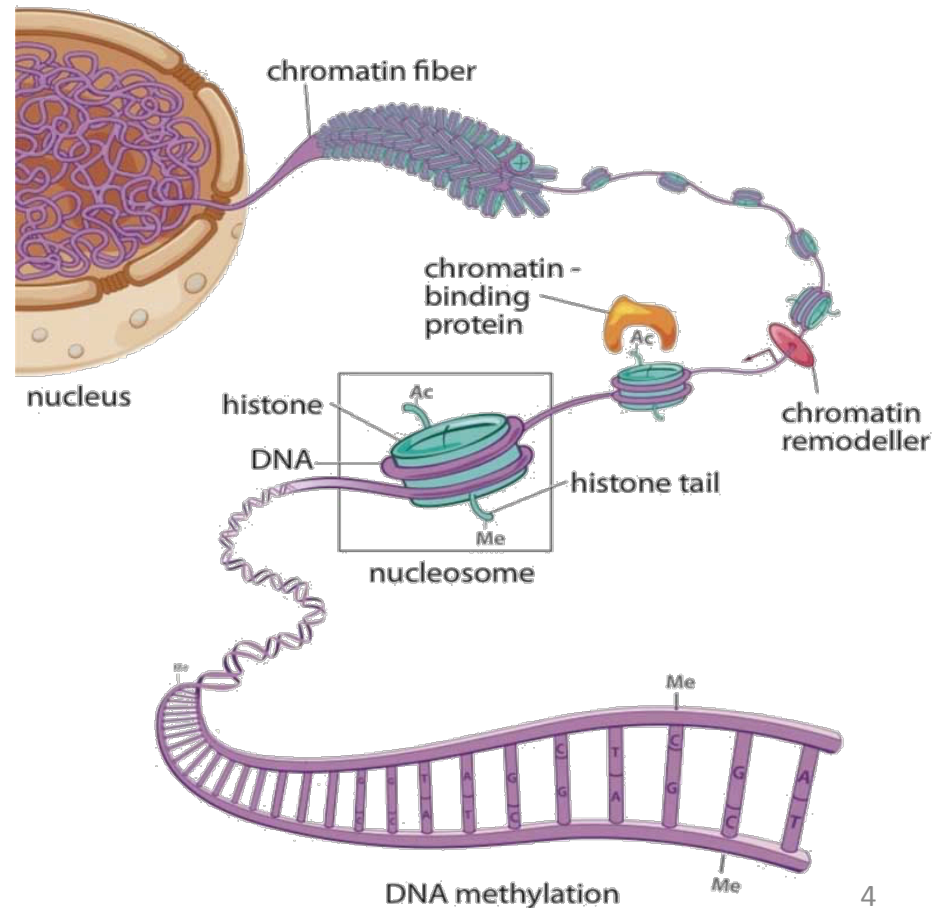
- All (almost) the cells of our body share the same genome but have very different gene expression programs....



# The code over the code

- The chromatin structure and the accessibility are mainly controlled by:

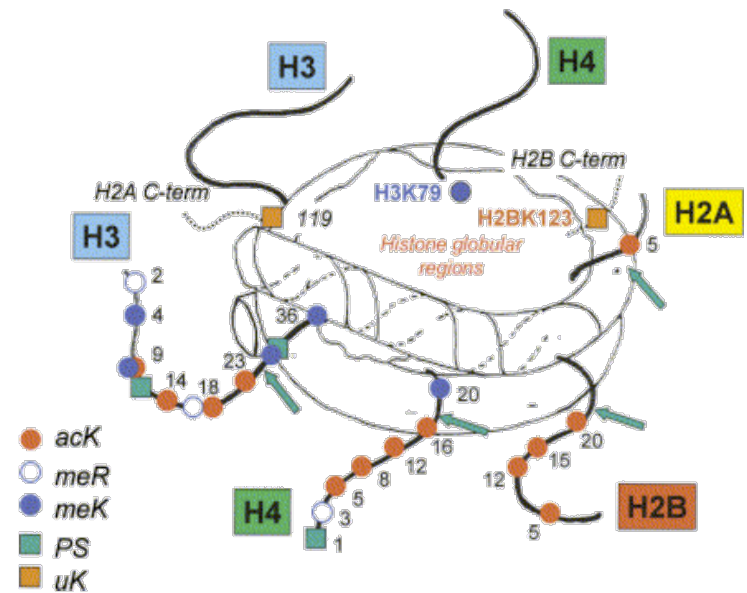
1. Nucleosome positioning,
2. DNA methylation,
3. Histone modifications.



# Histone Modifications

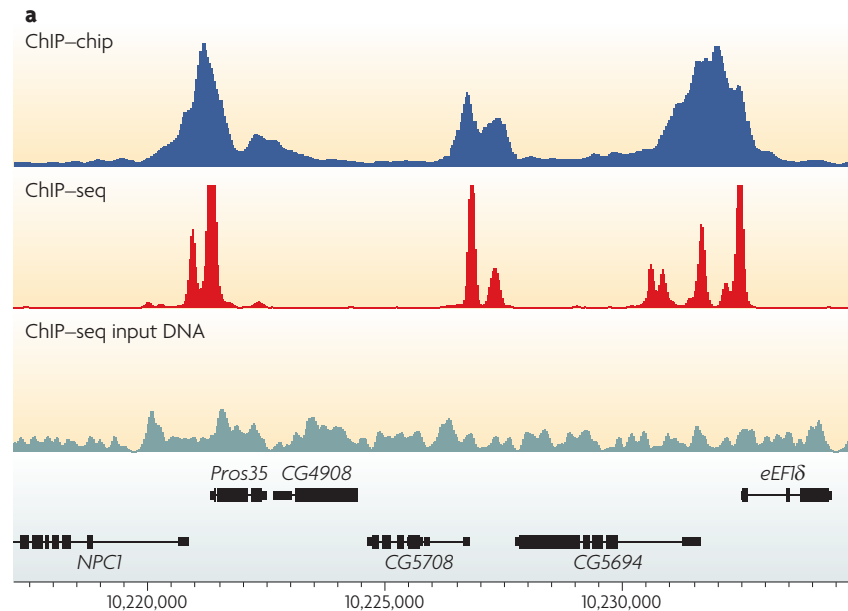
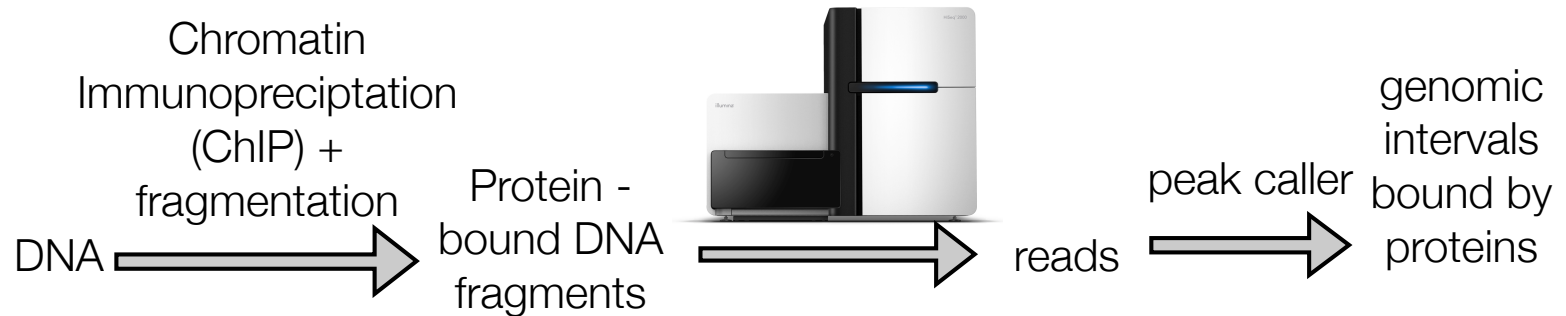
Specific histone modifications or combinations of modifications confer unique biological functions to the region of the genome associated with them:

- H3K4me3: *promoters, gene activation*
- H3K27me3: *promoters, poised enhancers, gene silencing*
- H2AZ: *promoters*
- H3K4me1: *enhancers*
- H3K36me3: *transcribed regions*
- H3K9me3: *gene silencing*
- H3k27ac: *active enhancers*

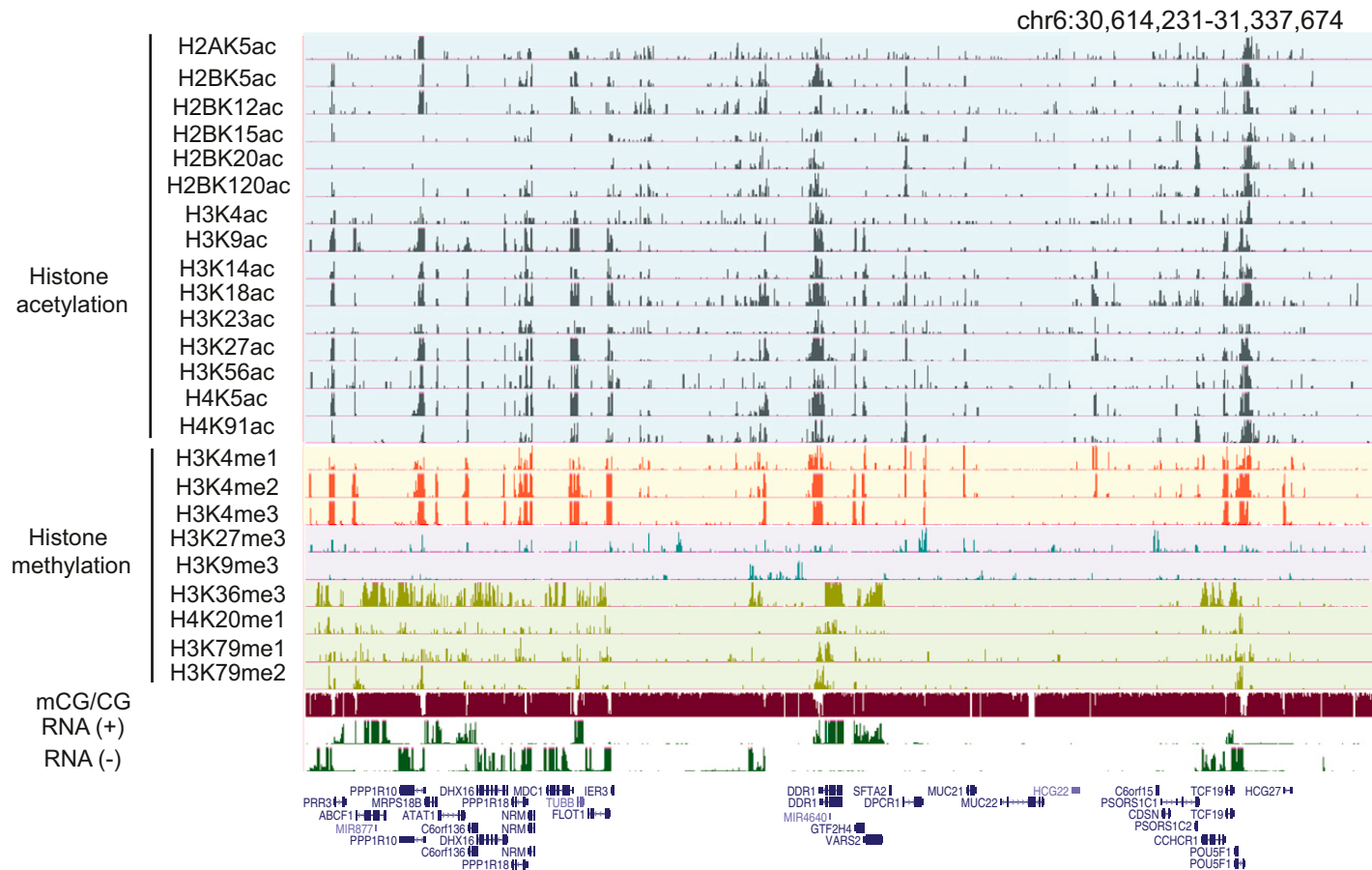


# ChIP-seq to measure histone data

Measuring the regulome (e.g., protein-binding of the genome)



# We can “call peaks” but...



Idea: We need a way to summarize the combinatorial patterns of multiple histone marks



# ChromHMM

“ChromHMM is a Java program for the learning and analysis chromatin states using a multivariate Hidden Markov Model that explicitly models the observed combination of marks”

<http://compbio.mit.edu/ChromHMM/>

## ARTICLE

doi:10.1038/nature09906

### Mapping and analysis of chromatin state dynamics in nine human cell types

Jason Ernst<sup>1,2</sup>, Pouya Kheradpour<sup>1,2</sup>, Tarjet S. Mikkelsen<sup>1</sup>, Noam Shores<sup>1</sup>, Lucas D. Ward<sup>1,2</sup>, Charles B. Epstein<sup>1</sup>, Xiaolan Zhang<sup>1</sup>, Li Wang<sup>1</sup>, Robbyn Issner<sup>1</sup>, Michael Coyne<sup>1</sup>, Manching Ku<sup>1,3,4</sup>, Timothy Durham<sup>1</sup>, Manolis Kellis<sup>1,3,4</sup> & Bradley E. Bernstein<sup>1,3,4</sup>



# ChromHMM and Chromatin States

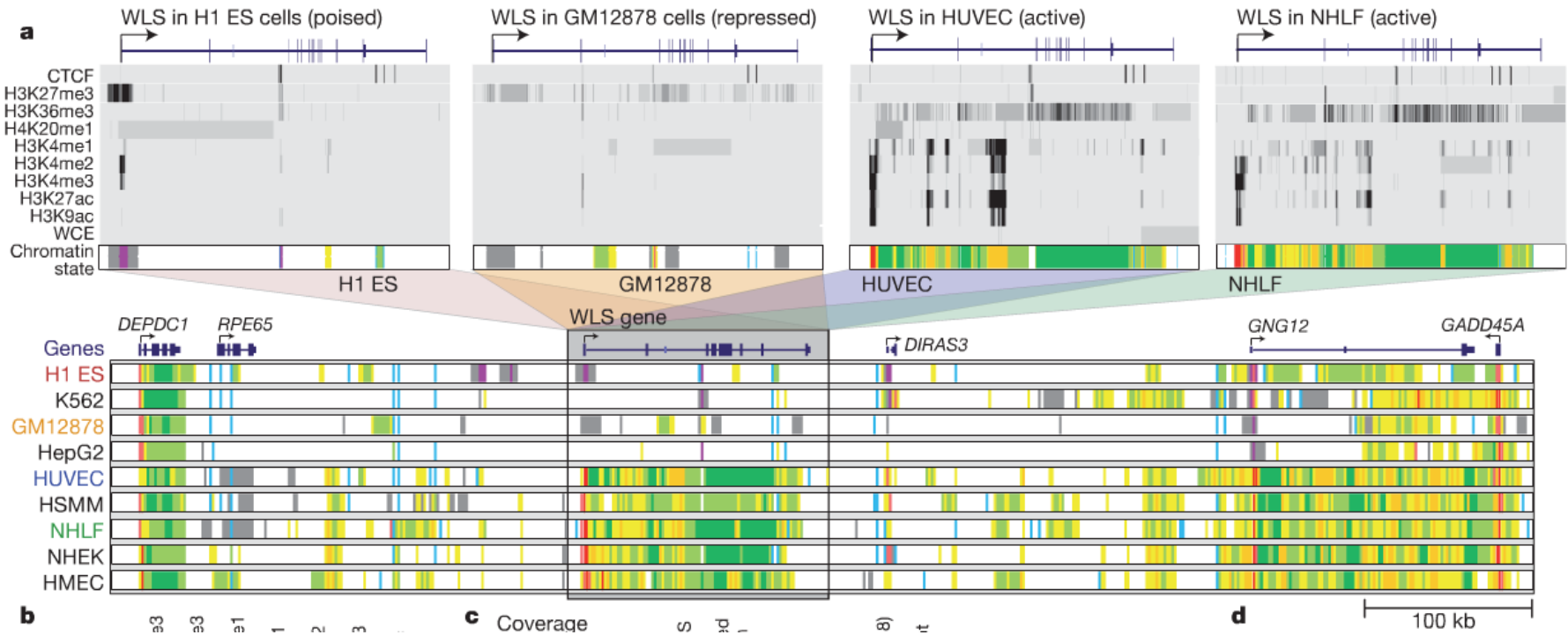
- **Chromatin states** are defined based on different combinations of histone modifications and correspond to different functional regions

Chromatin States	State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Candidate state annotation
	1	16	2	2	6	17	93	99	96	98	2	
2	12	2	6	9	53	94	95	14	44	1	Weak Promoter	
3	13	72	0	9	48	78	49	1	10	1	Inactive/poised Promoter	
4	11	1	15	11	96	99	75	97	86	4	Strong enhancer	
5	5	0	10	3	88	57	5	84	25	1	Strong enhancer	
6	7	1	1	3	58	75	8	6	5	1	Weak/poised enhancer	
7	2	1	2	1	56	3	0	6	2	1	Weak/poised enhancer	
8	92	2	1	3	6	3	0	0	1	1	Insulator	
9	5	0	43	43	37	11	2	9	4	1	Transcriptional transition	
10	1	0	47	3	0	0	0	0	0	1	Transcriptional elongation	
11	0	0	3	2	0	0	0	0	0	0	Weak transcribed	
12	1	27	0	2	0	0	0	0	0	0	Polycomb-repressed	
13	0	0	0	0	0	0	0	0	0	0	Heterochrom; low signal	
14	22	28	19	41	6	5	26	5	13	37	Repetitive/CNV	
15	85	85	91	88	76	77	91	73	85	78	Repetitive/CNV	

Chromatin Mark Observation Frequency (%)

- The goal is to segment the genome into biologically meaningful units.

# ChromHMM and segmentation



# ChromHMM in practice: gather the ingredients

- Required:
  1. Java virtual machine (<http://java.com/>)
  2. ChromHMM software (<http://compbio.mit.edu/ChromHMM/ChromHMM.zip>)
  3. Aligned ChIP-seq files for different histone modifications for example from the ENCODE portal (<https://www.encodeproject.org/>)
- Optionally, if we want to use it on your data:
  1. Raw or aligned reads for different histone modifications
  2. A fast aligner aligner like Bowtie (<http://bowtie-bio.sourceforge.net/bowtie2>) or BWA (<http://bio-bwa.sourceforge.net/>)
  3. Bedtools (<https://github.com/arq5x/bedtools2>)

# The Workflow

1. Get ChIP-seq raw reads for different histone modifications
2. Align the reads to a reference genome
3. Convert aligned reads in bed format
4. Create Binned and Binarized Tracks
5. Train the model
6. Infer the states
7. Interpretation

# Align the reads

- Starting from a file containing raw reads (usually a fastq file) you need to align them to a reference genome to get a .bam file (binary aligned file). You can use Bowtie or BWA (links in slide #11).



- Or you can download many aligned samples from the encode portal <https://www.encodeproject.org/>

# Convert aligned reads to bed format

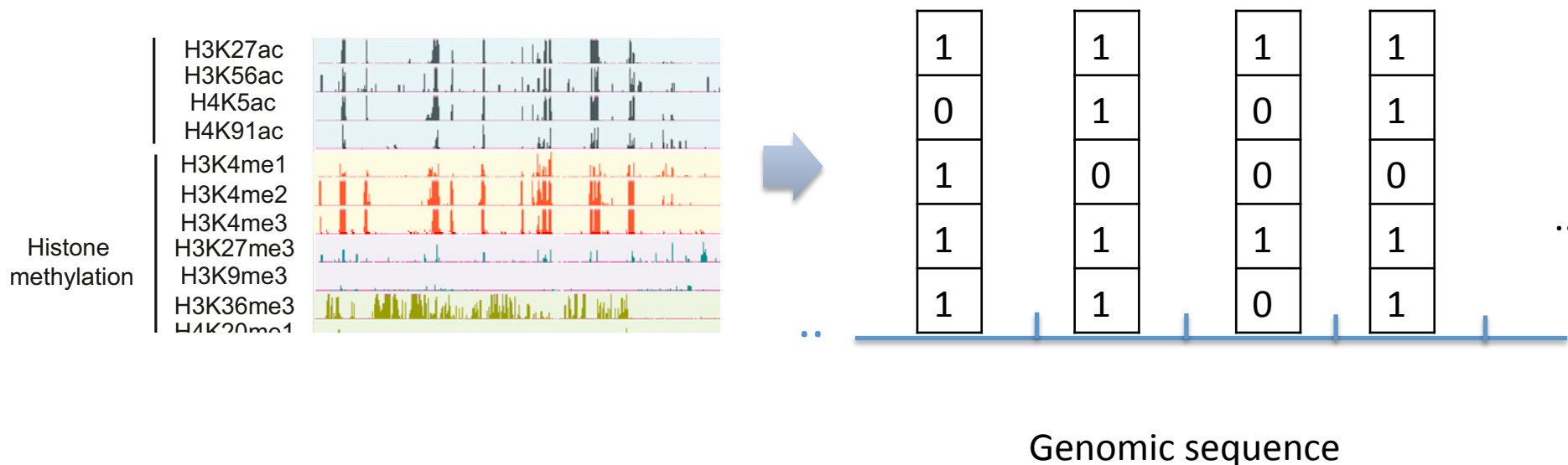
- ChromHMM needs the aligned reads in .bed format



```
bedtools bamtobed -i cell1_mark1.bam > ~/data/cell1_mark1.bed
```

# Create Binned and Binarized Tracks

- ChromHMM quantify the presence or absence of each mark in bins of fixed size





# Create Binned and Binarized Tracks

- ```
java -mx4000M -jar ChromHMM.jar  
BinarizeBed -b 200 CHROMSIZES/hg18 ~/  
data/ cellmarkfiletable.txt  
SAMPLEDATA_HG18
```
- Inside the cellmarkfiletable.txt:

```
cell11 mark1 cell11_mark1.bed cell11_control.bed  
cell11 mark2 cell11_mark2.bed cell11_control.bed  
cell12 mark1 cell12_mark1.bed cell12_control.bed  
cell12 mark2 cell12_mark2.bed cell12_control.bed
```

# Train the model and segment the genome



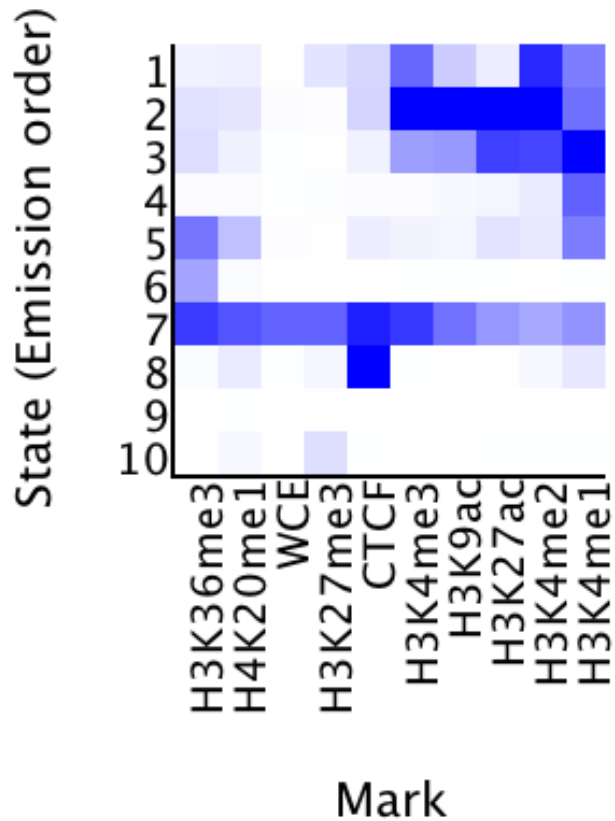
```
java -mx1600M -jar ChromHMM.jar LearnModel  
SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18
```

# Output of ChromHMM

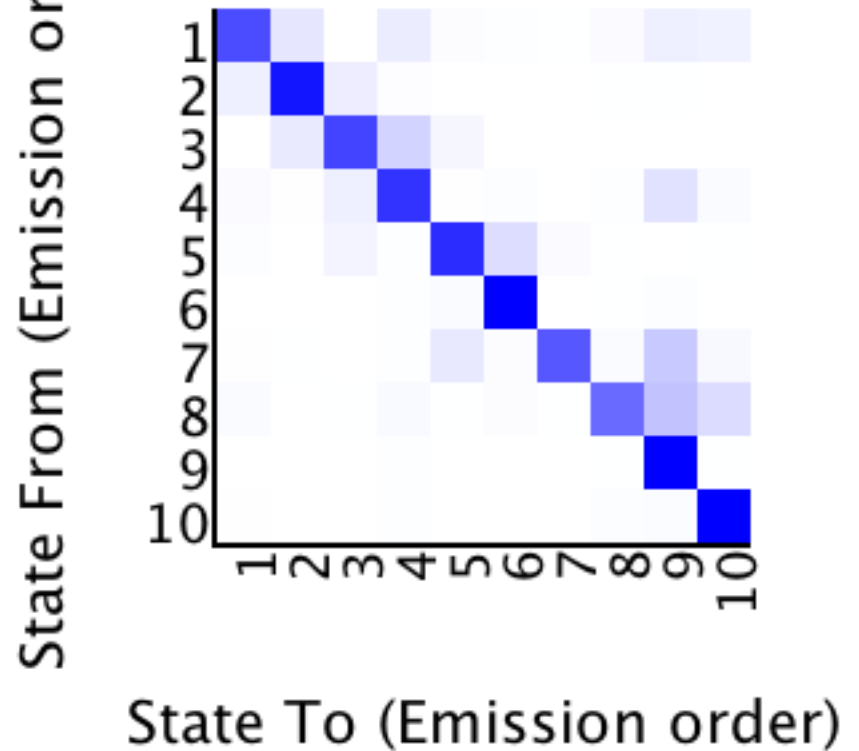
- ChromHMM generates a nice HTML report called **webpage\_N.html** (N is the number of states used) with many useful information :
  1. Model learned: transition and emission parameters
  2. Enriched functional categories
  3. Bed files to visualize the segmentation

# Transition and emission Parameters

Emission Parameters

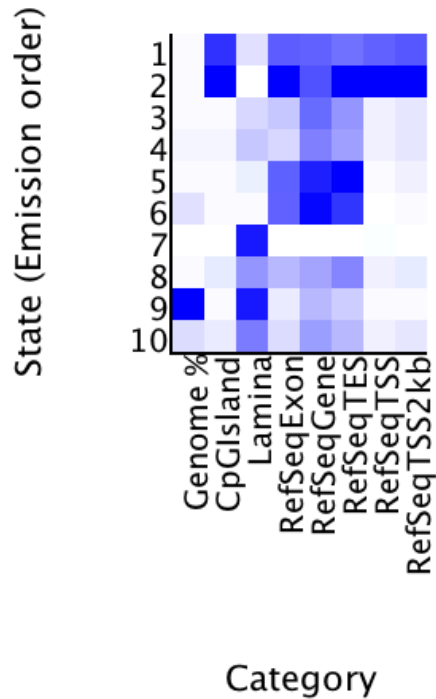


Transition Parameters

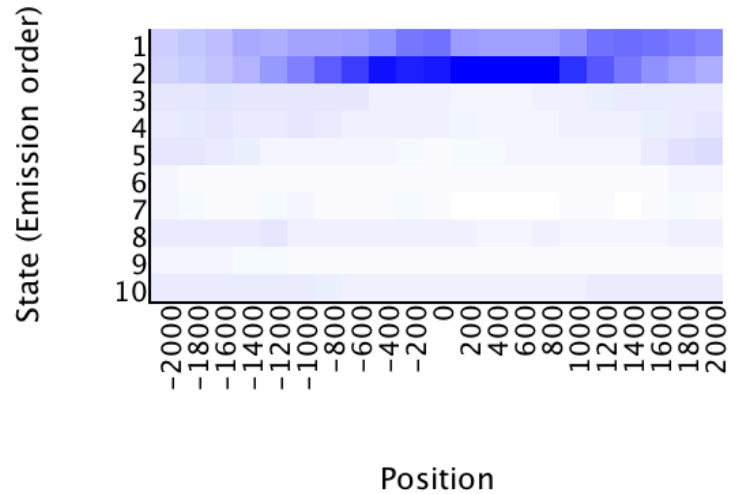


# Enriched functional category

Fold Enrichment GM12878\_10

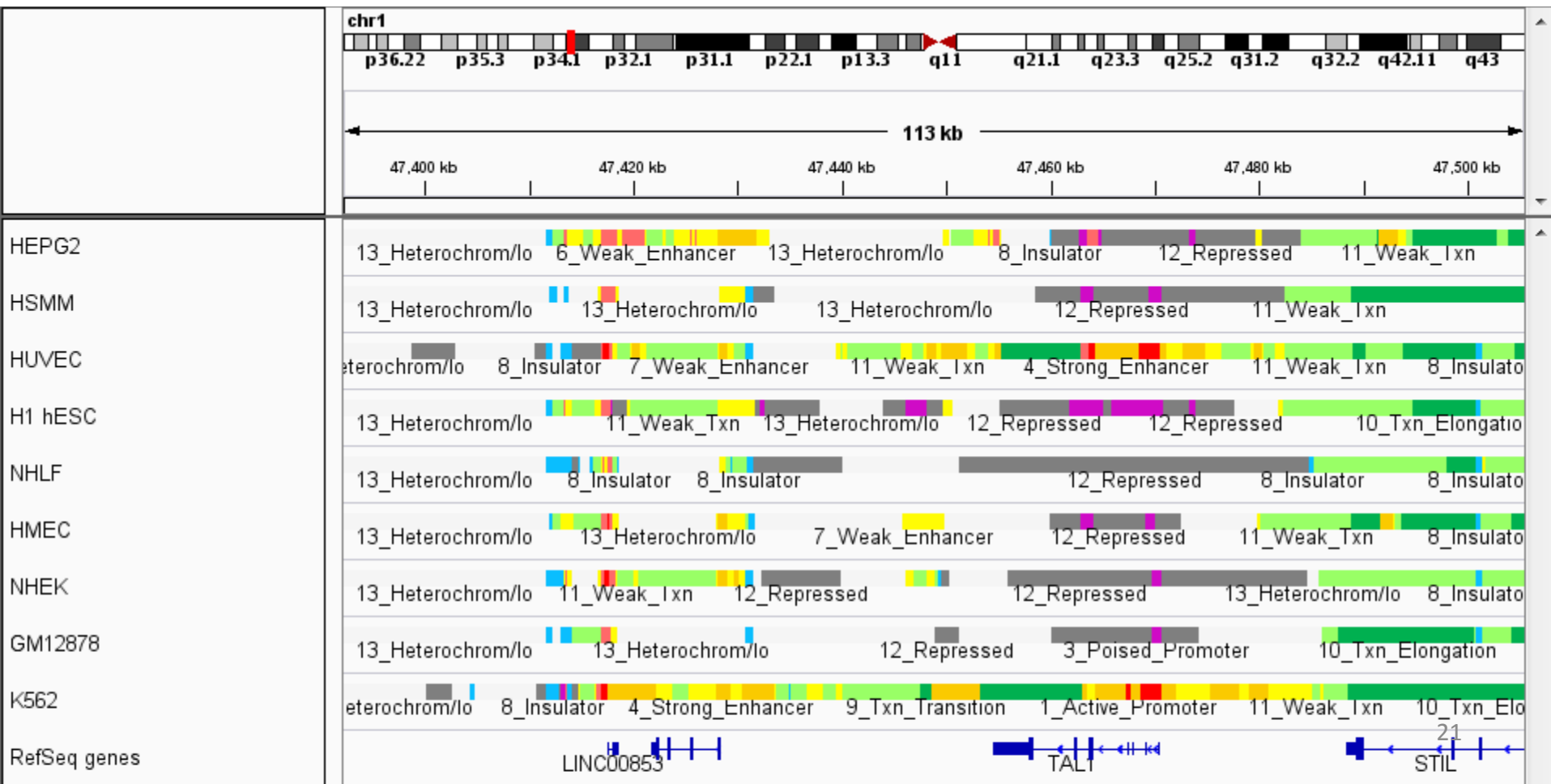


Fold Enrichment GM12878\_10 RefSeqTSS



# Visualize the segmentation

- Genome Browser: <https://genome.ucsc.edu/>
- IGV: <https://www.broadinstitute.org/igv/>



# Further References

- Other models are available to segment the genome in chromatin states:
  1. Segway:  
<http://pmgenomics.ca/hoffmanlab/proj/segway/>
  2. Spectacle: <https://github.com/jiminsong/Spectacle>
  3. DI-HMM (soon available) GC Yuan/M Kellis



# Questions?

