Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 05

Pierre-Luc Germain

ETH Zürich

Plan

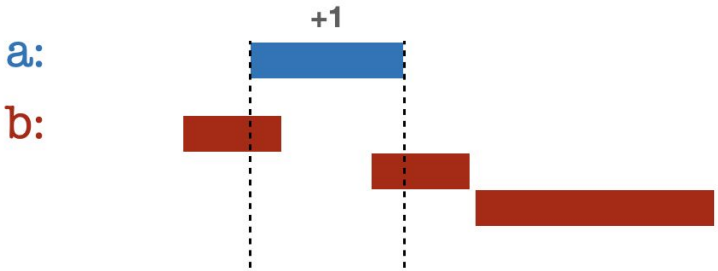
- Debriefing on the assignment
- The ‘histone code’ & functional elements
- More on overlaps and comparing signals

Debriefing on the assignments

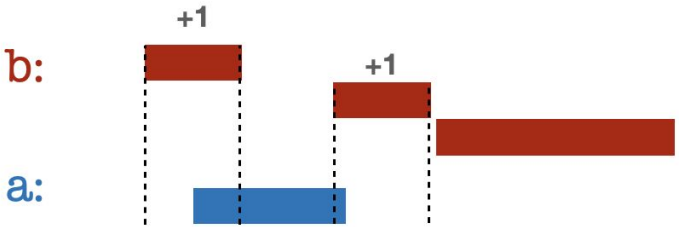
Symmetry of overlaps:

a: query
b: subject

overlapsAny(a, b)
findoverlaps(a, b)



sum(overlapsAny(a, b))=1



sum(overlapsAny(b, a))=2

findOverlaps **vs** overlapsAny

```
ov <- findOverlaps(p300, h3k4me3)
ov
```

```
## Hits object with 634 hits and 0 metadata columns:
##         queryHits subjectHits
##         <integer>   <integer>
##     [1]         1         678
##     [2]         2         556
##     [3]         5        1681
##     [4]         9        1990
##     [5]        12         547
##     ...       ...         ...
##   [630]      1794         572
##   [631]      1795        1782
##   [632]      1798         423
##   [633]      1800        2196
##   [634]      1800        2197
##   -------
##   queryLength: 1801 / subjectLength: 2865
```

```
# like this we count p300 peaks which overlap more than one H3K4me3 peak several times
length(queryHits(ov))/length(p300)
```

```
## [1] 0.3520267
```

findOverlaps **vs** overlapsAny

```
ov <- findOverlaps(p300, h3k4me3)
ov
```

```
## Hits object with 634 hits and 0 metadata columns:
##         queryHits subjectHits
##         <integer>   <integer>
##     [1]         1         678
##     [2]         2         556
##     [3]         5        1681
##     [4]         9        1990
##     [5]        12         547
##     ...       ...         ...
##   [630]      1794         572
##   [631]      1795        1782
##   [632]      1798         423
##   [633]      1800        2196
##   [634]      1800        2197
##   -------
##   queryLength: 1801 / subjectLength: 2865
```

be aware of that one peak in the query might overlap several in the subjects

```
# like this we count p300 peaks which overlap more than one H3K4me3 peak several times
length(queryHits(ov))/length(p300)
```

```
## [1] 0.3520267
```

`findOverlaps` **vs** `overlapsAny`

correct solution

```
# need to use unique
length(unique(queryHits(ov)))/length(p300)
```

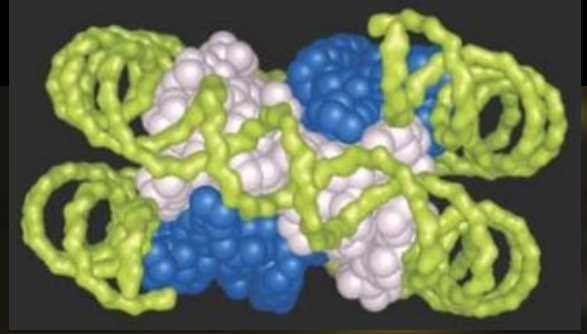
```
## [1] 0.3053859
```

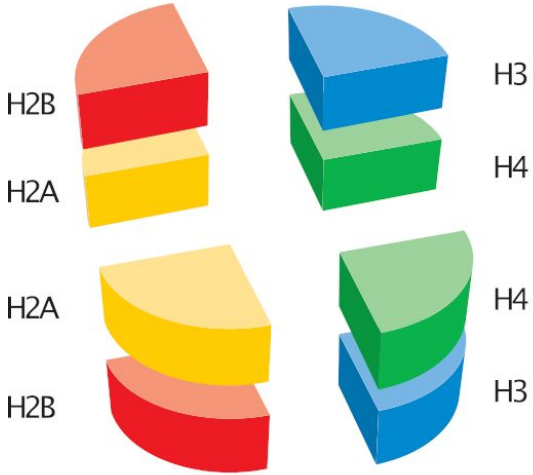
```
sum(overlapsAny(p300, h3k4me3))/length(p300)
```

```
## [1] 0.3053859
```

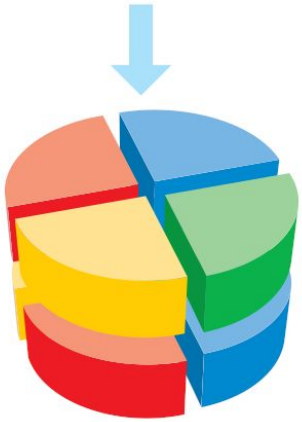
Nucleosome

(Krebs et al., 2018)



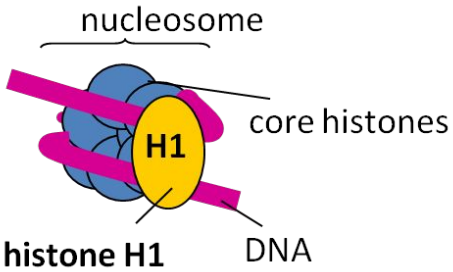


core histones



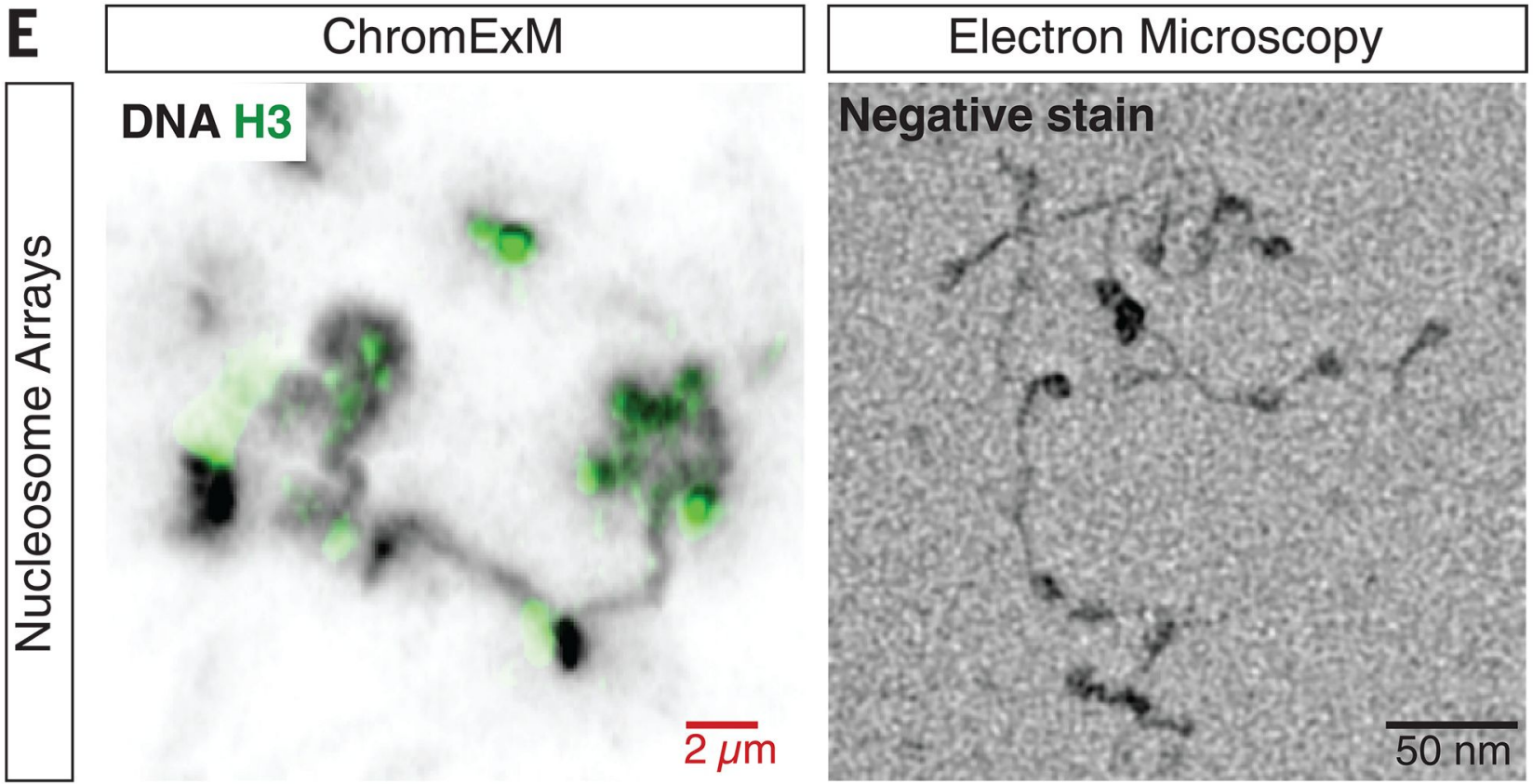
histone octamer

beads–on–a–string nucleosome array



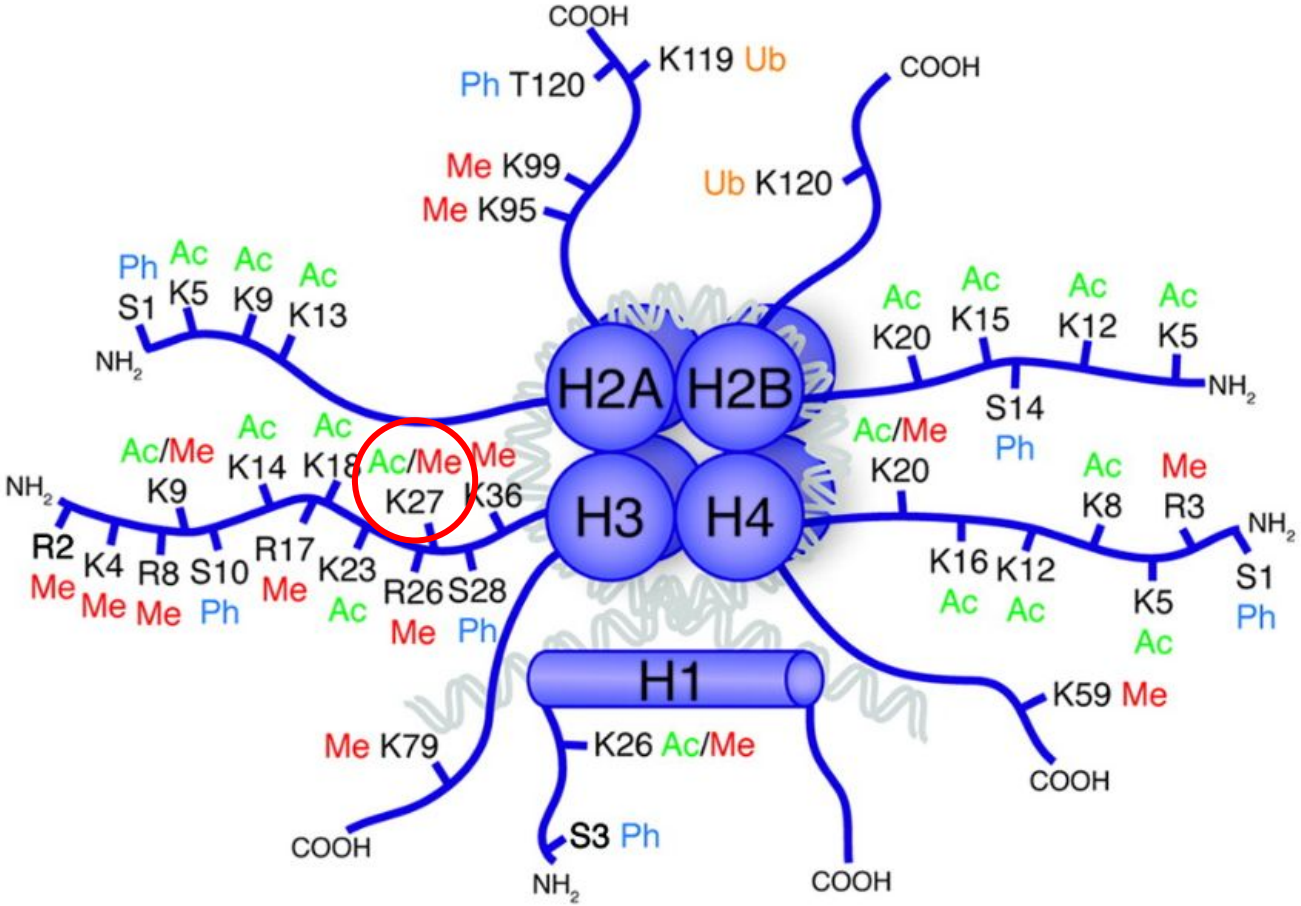
(Adapted from David O Morgan - The Cell Cycle. Principles of Control. Wikimedia Commons)

Viewing the beads-on-a-string of DNA and nucleosomes



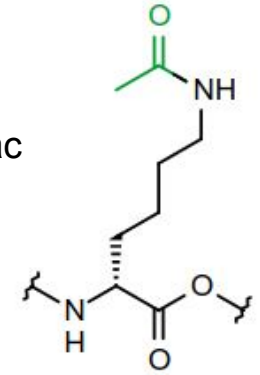
(Adapted from Pownall et al., Science 2023)

Many residues on the histone tails can be post-translationally modified

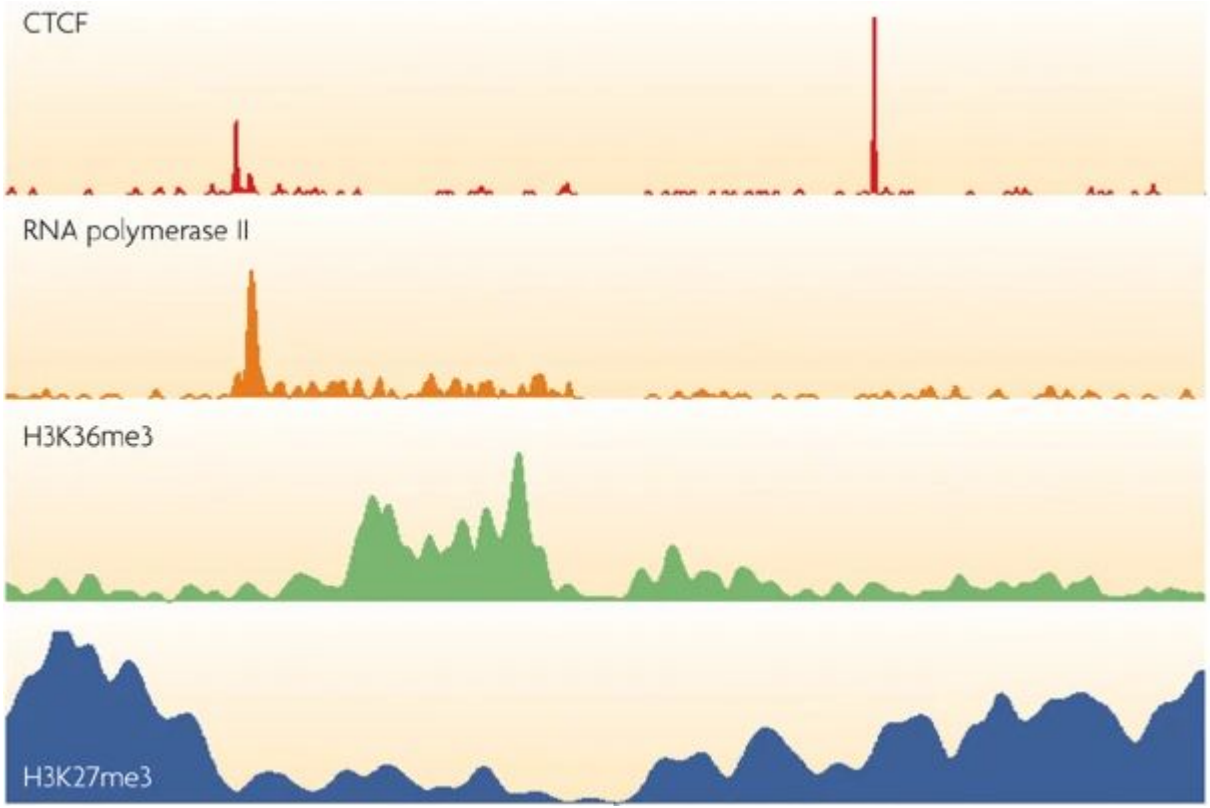


(Adapted from Tollervey & Lunyak, Epigenetics 2012)

H3K27ac



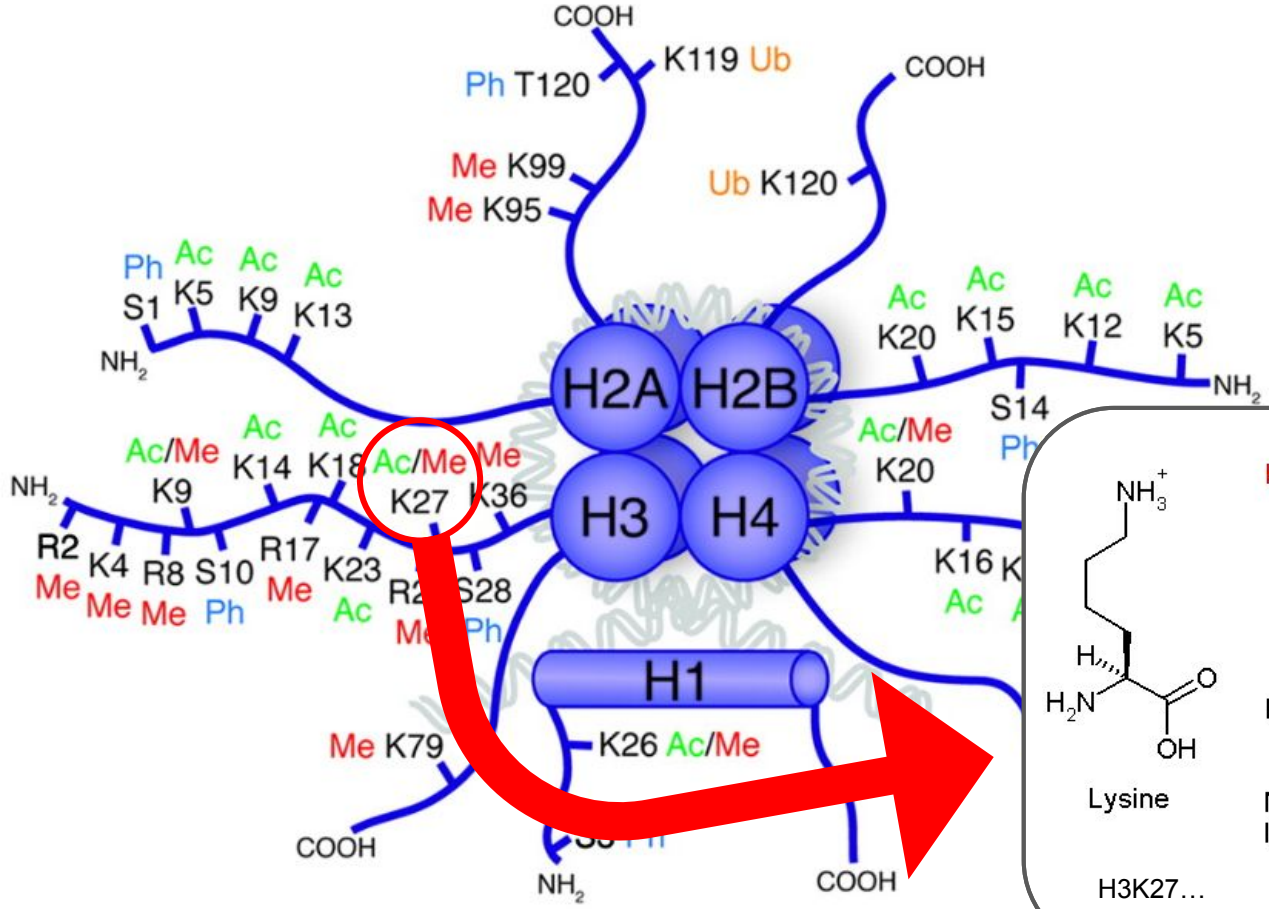
Some histone modifications appear to be very localized, e.g. happening on a specific nucleosome, while most are much more broadly distributed



The strategy of calling ‘peaks’ must therefore be adapted
(e.g. “broad” option of most peak-callers)

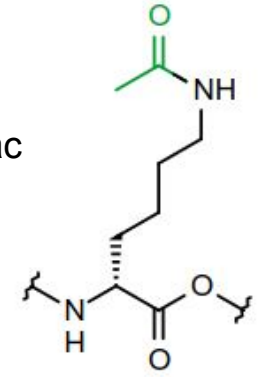
(Park, Nat Rev Gen 2009)

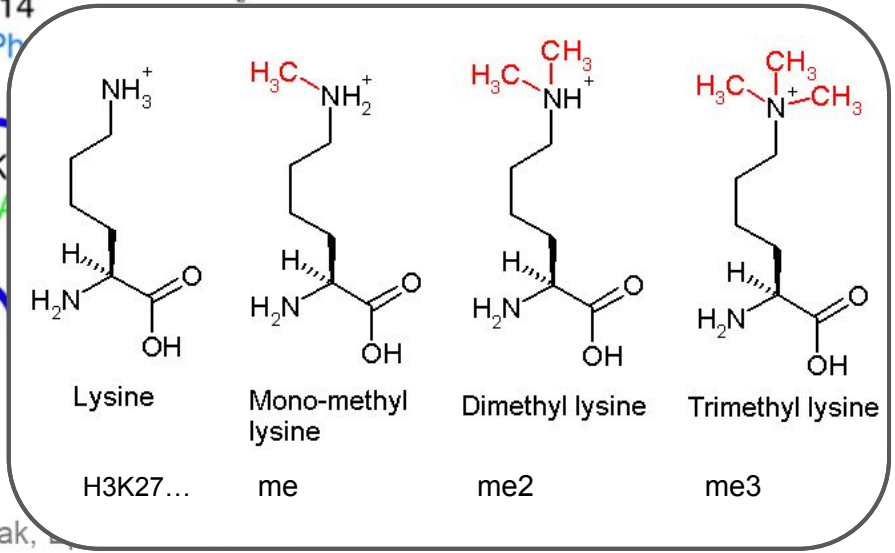
Many residues on the histone tails can be post-translationally modified



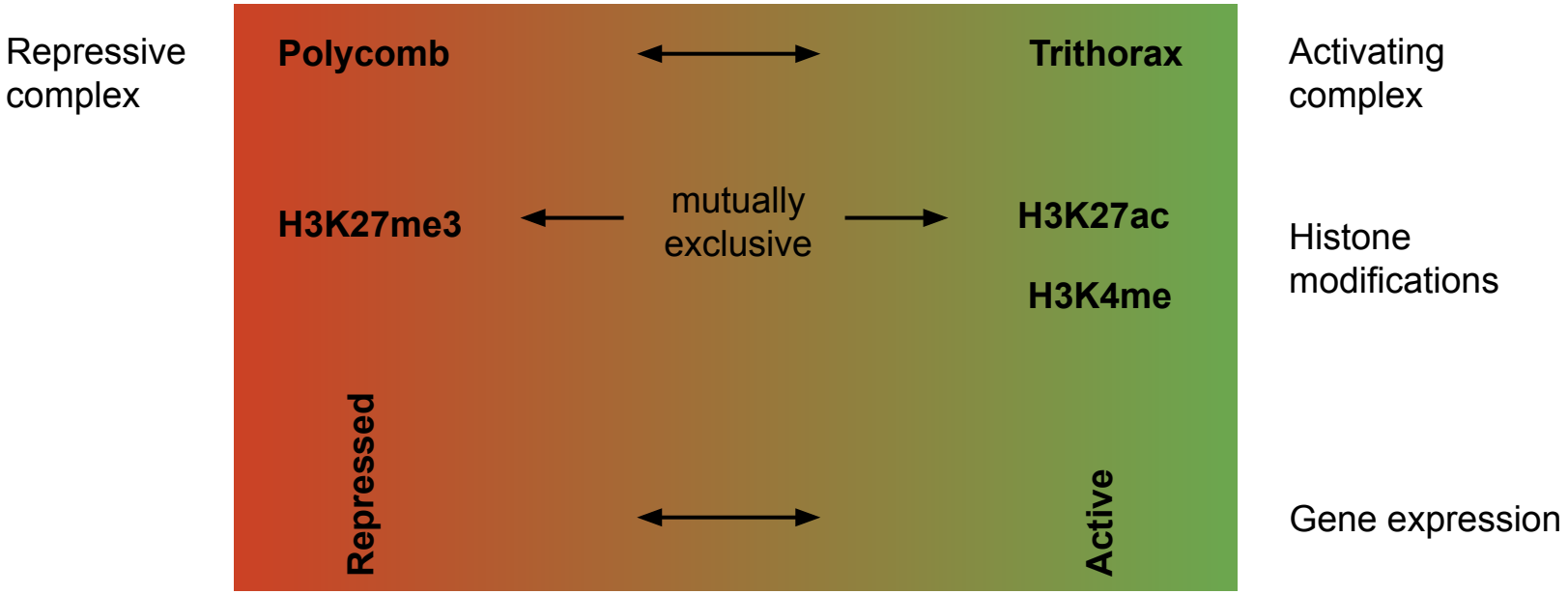
(Adapted from Tollervey & Lunyak,

H3K27ac





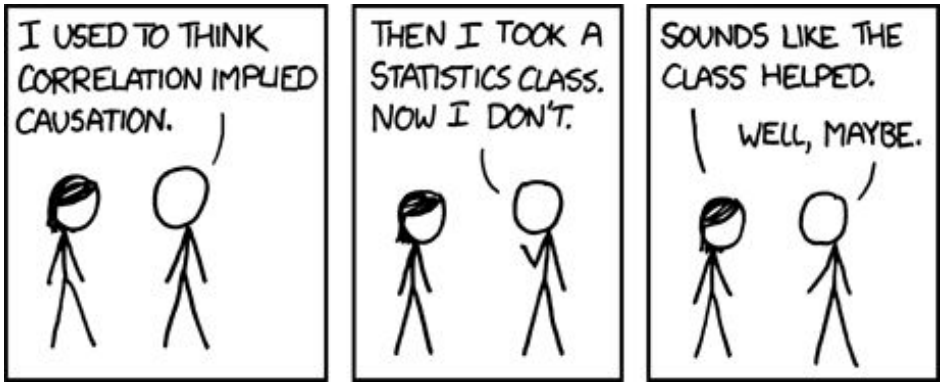
There is a very strong association of certain histone marks and activation or repression



But which comes first?

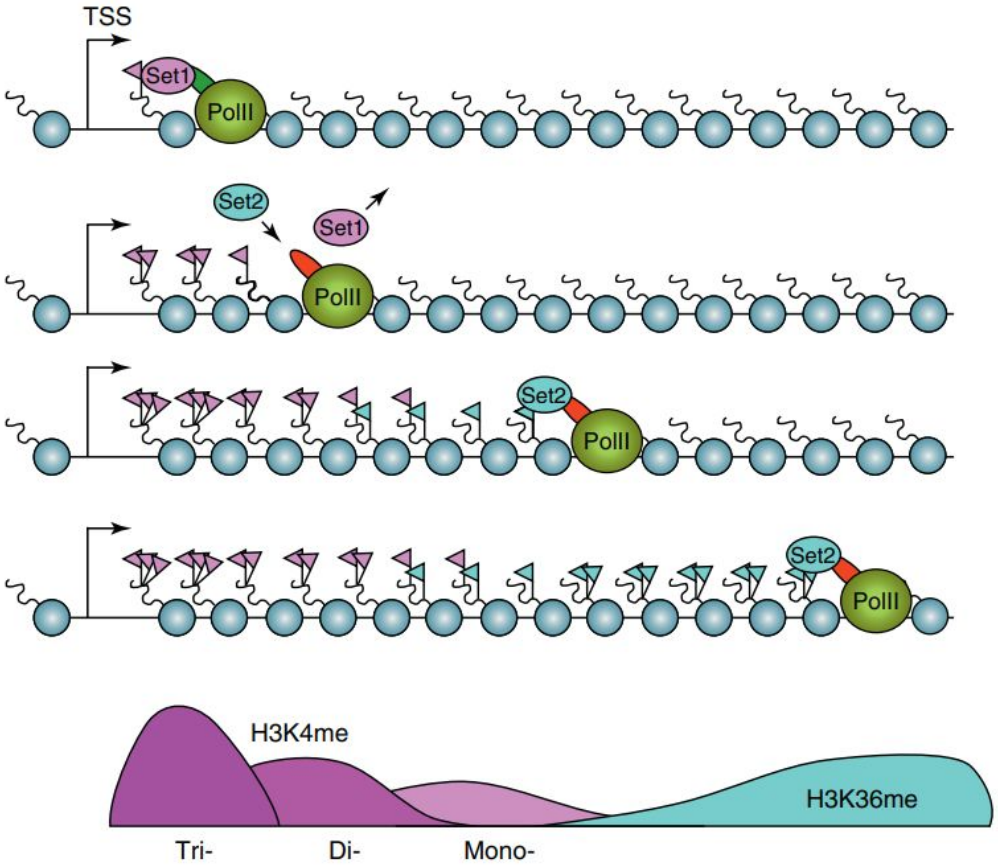
Causality or correlation?

Are histone modifications **responsible** for activation/repression, or are they merely associated **side-effects**?



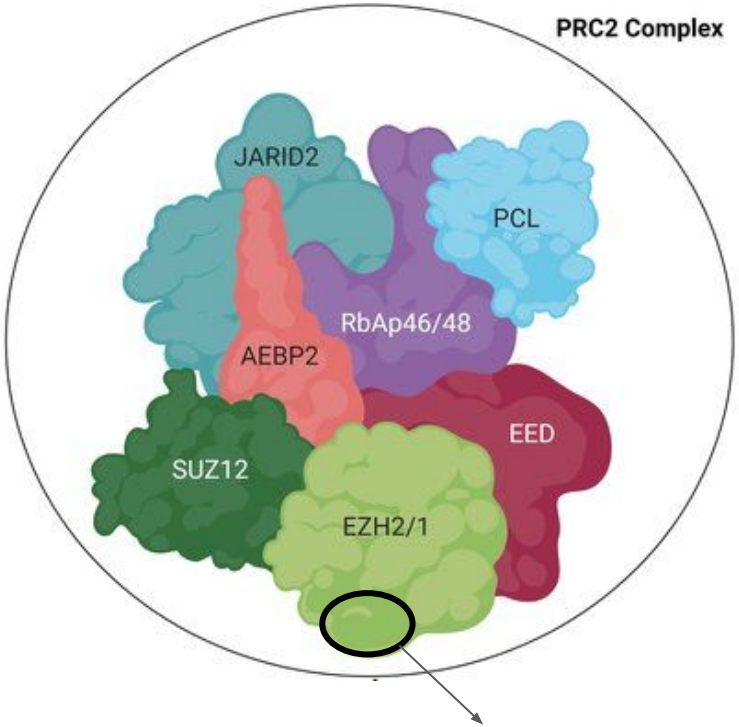
(https://xkcd.com/552)

Transcription-mediated histone modification



(Henikoff and Shilatifard 2011)

The example of H3K27me3, chiefly deposited by the polycomb repressive complex (PRC2)



EHZ2's SET domain catalyzes the addition of a 3rd methyl group to H3K27, i.e. H3K27me2 → H3K27me3

(Adapted from Zoroddu et al., 2021)

Abolishing the enzymatic activity of *Ezh2*, the gene responsible for depositing H3K27me3, abolishes (most of) the mark but does not prevent the repression of the target genes, nor cellular reprogramming

(Fragola et al., PLoS Genetics 2013)

Causality or correlation?

Other examples…

For H3K27ac, blocking the deposition of the mark does not prevent pioneering factors (PF) from opening local chromatin, whereas blocking the factors prevents H3K27ac to most sites

(Miao et al., Molecular Cell 2022)

Similarly, the loss of H3K4me3 appears to have no effect on nascent transcription

(Murray et al., bioRxiv 2019)

nature

Article | Open Access | Published: 01 March 2023

H3K4me3 regulates RNA polymerase II promoter-proximal pause-release

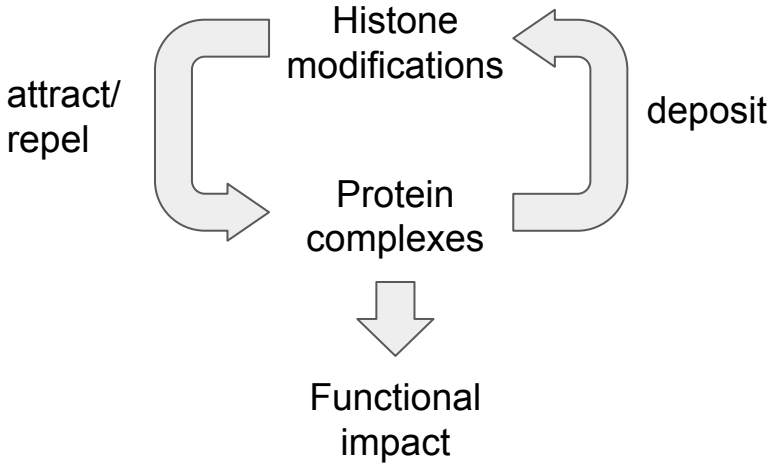
Hua Wang, Zheng Fan, Pavel V. Shliaha, Matthew Miele, Ronald C. Hendrickson, Xuejun Jiang & Kristian Helin

Nature **615**, 339–348 (2023)

"acute **loss of H3K4me3 does not have detectable effects on transcriptional initiation** but leads to a widespread decrease in transcriptional output, an increase in RNA polymerase II (RNAPII) pausing and slower elongation. We show that H3K4me3 is required for the recruitment of the integrator complex subunit 11 (INTS11), which is **essential for the eviction of paused RNAPII and transcriptional elongation**."

Causality or correlation?

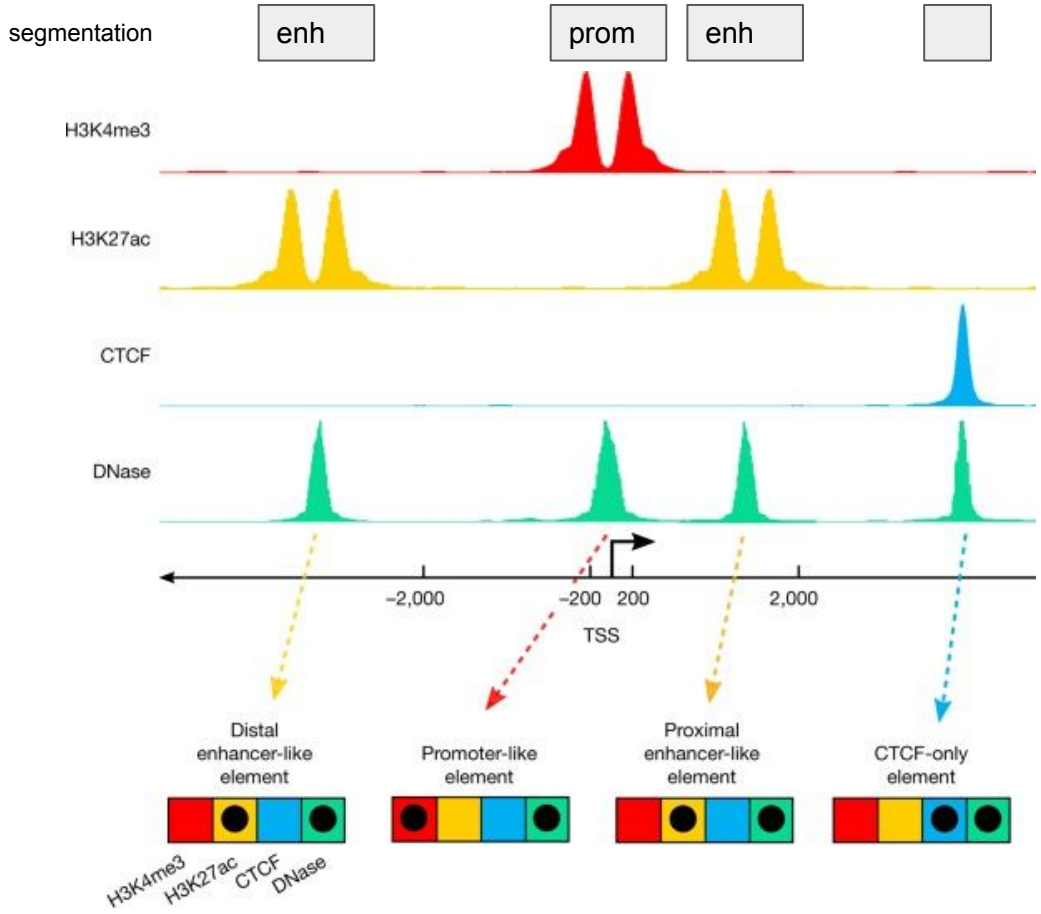
Most likely somewhere in the middle, depending on the modification/context



Whether they’re causative or not, they can serve as **proxies** for function.

This means that profiling a few histone modifications gives an overview of the epigenomic landscape of a cellular state which would otherwise require profiling all the potentially-relevant factors/complexes

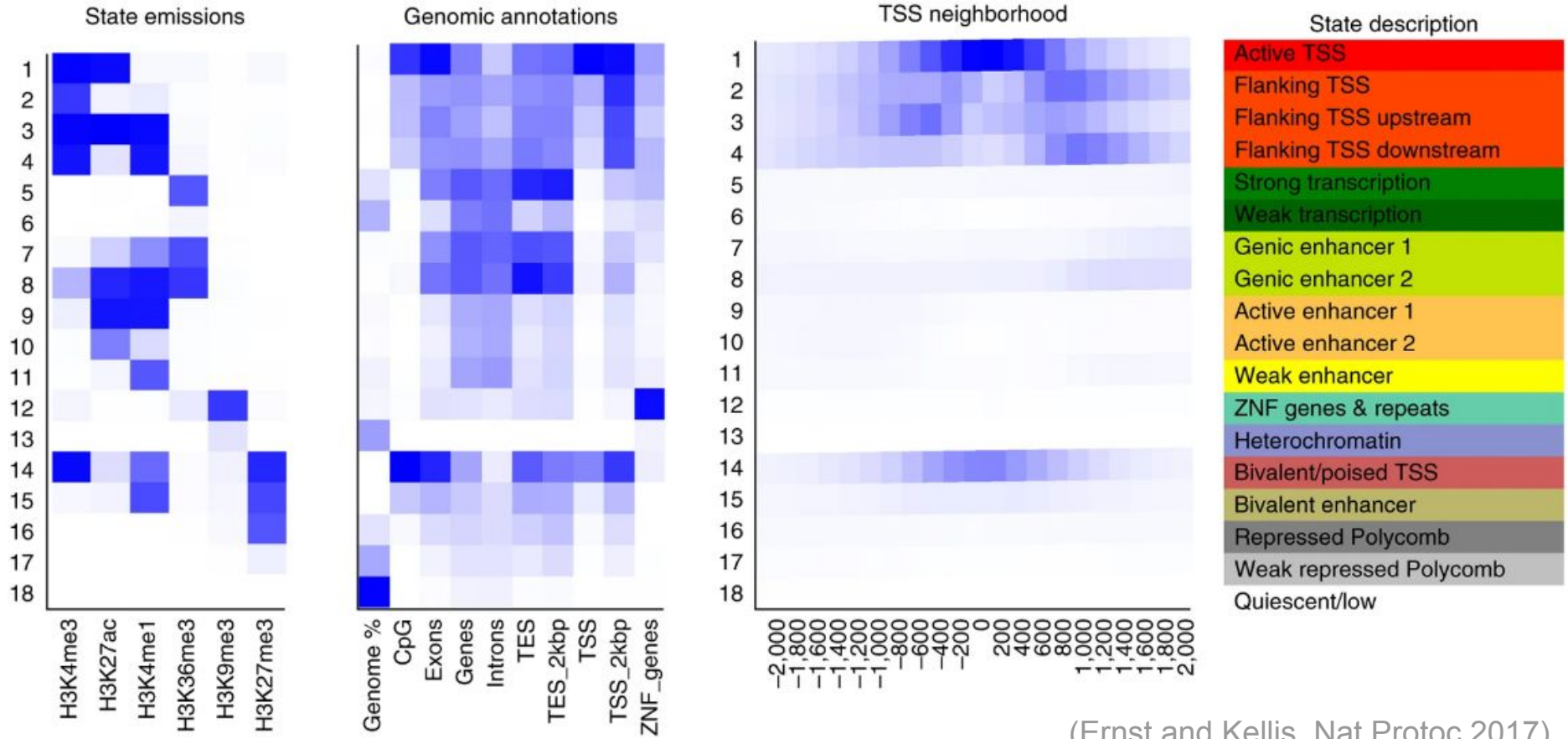
A signature-based encyclopedia of DNA elements



ENCODE's "signature strategy":

- Different types of functional genetic elements are associated with different chemical signatures
- We can identify functional elements by identifying these signatures genome-wide

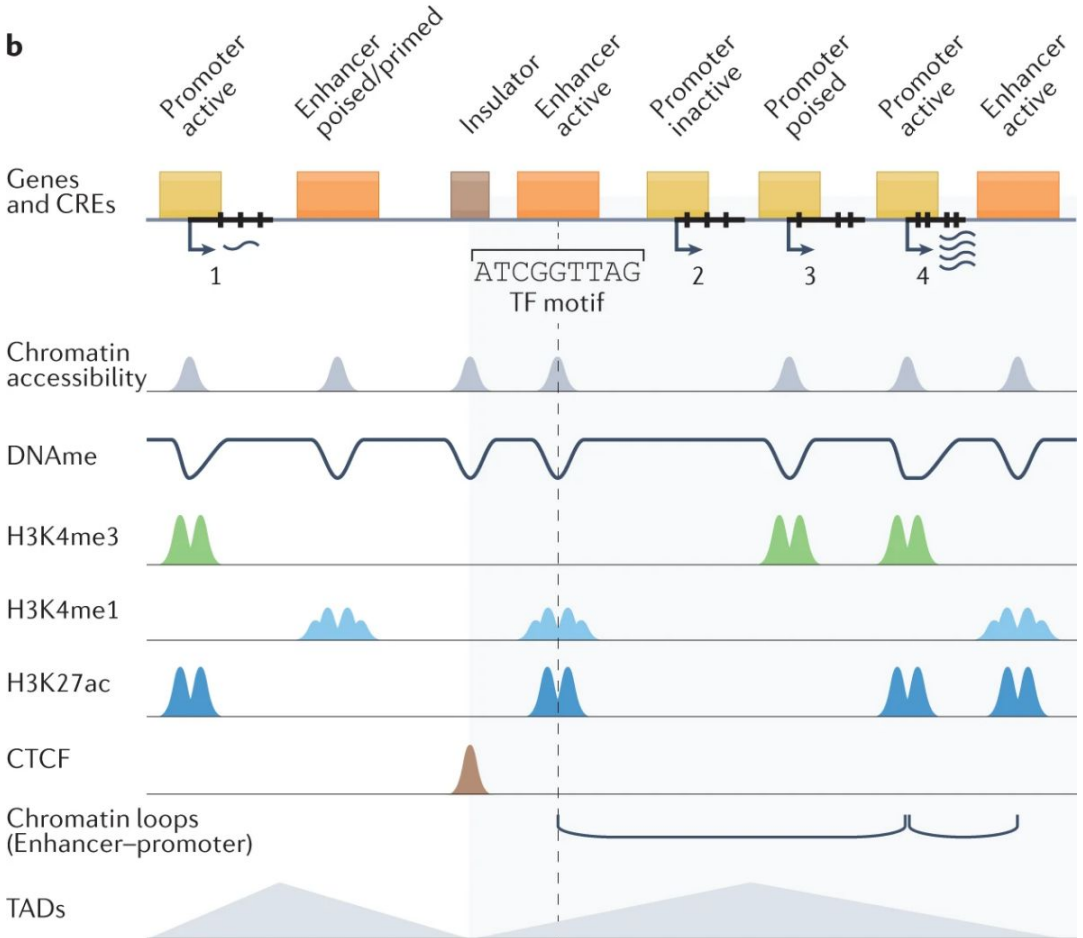
So how many kinds of functional elements/states are there?



(Ernst and Kellis, Nat Protoc 2017)

So how many kinds of functional elements/states are there?

More recent papers/reviews tend to narrow down to a very basic set of elements…



(Preissl, Gaulton & Ren, Nat Rev Gen 2023)

Some stuff is pretty clear:

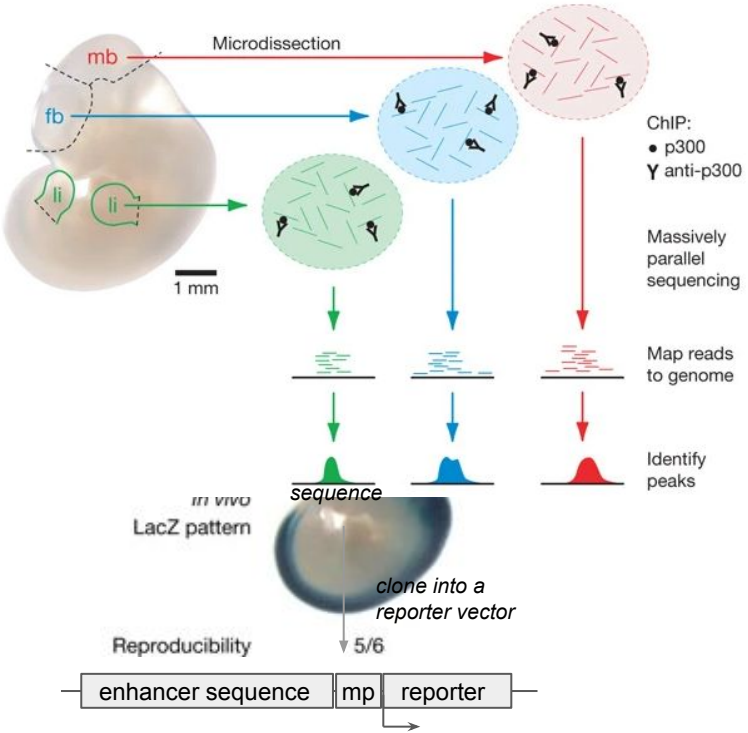
- **Transcription start site (TSS):**
  - **H3K4me3** is almost always associated with active/poised TSS
  - Active TSS are marked by **H3K27ac**
  - So-called “poised” (or bivalent) TSS are instead marked by both **H3K4me3** and **H3K27me3**

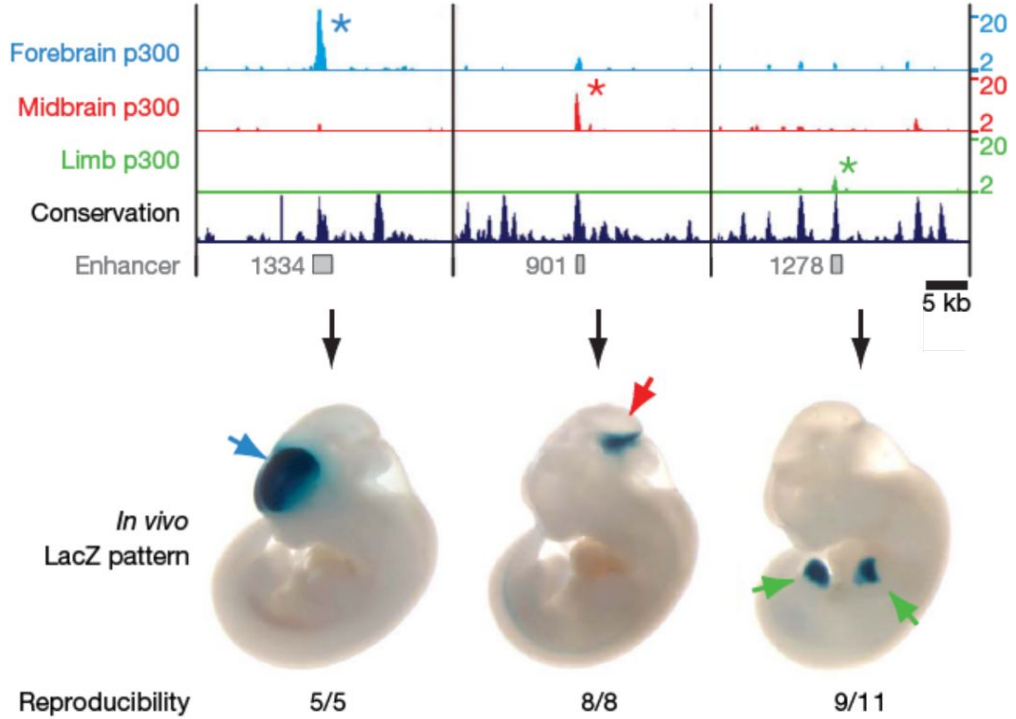
- **Enhancers:**
  - Most enhancers have **H3K4me1**
  - Active enhancers are marked by **H3K27ac**
  - So-called “poised” (or bivalent) enhancers are marked by **H3K4me1** and **H3K27me3**

- Repressed elements are marked by **H3K27me3**

- Heterochromatin is marked by **H3K9me3**
- Insulators: CTCF (+cohesin etc.)

p300 and validation of enhancer activity





(Adapted from Visel et al., 2009)

Vista enhancers: http://enhancer.lbl.gov

Assignment

- Using the peaks you downloaded last week, identify bivalent domains (H3K27me3 + H3K4me3) in human embryonic stem cells (ESC)

- What happens to those regions upon differentiation?
  - Choose **one** differentiated cell type (e.g. erythroblasts, fibroblast, B cells, etc.)
  - Download the H3K27me3 and H3K4me3 **peaks** from this cell type
  - How many of the ESC bivalent domains are, in this differentiated cell type, overlapping either mark or their combination (in this differentiated cell type)?

- Don’t forget to upload your assignment as “`assignment.html`” !