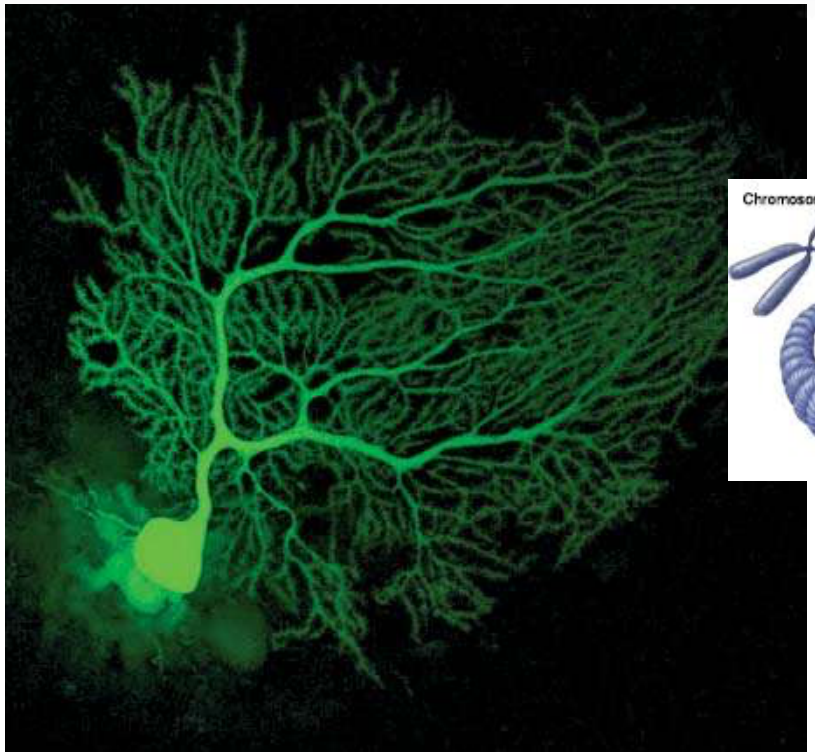
Computational Epigenetics

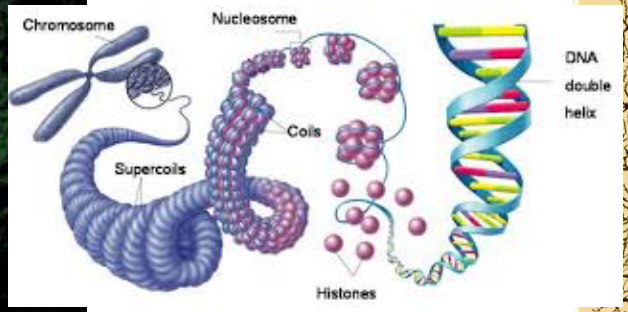
Gabriele Schweikert
University of Dundee, UK
Universität Tübingen, Germany

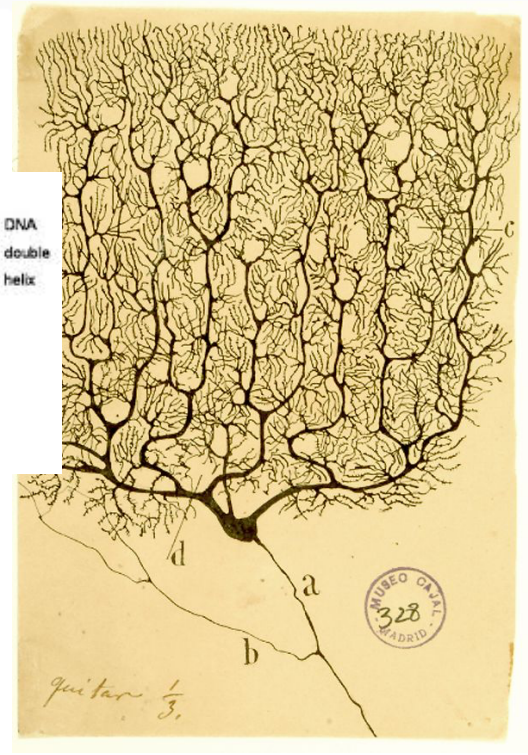
Take home message:
Bioinformatics is a little bit like learning to walk on a slack line: You will fall and encounter errors all the time. This is not because you are not clever enough. Falling is all part the fun. You just get up and try again.

Purkinje cells



Genetic code

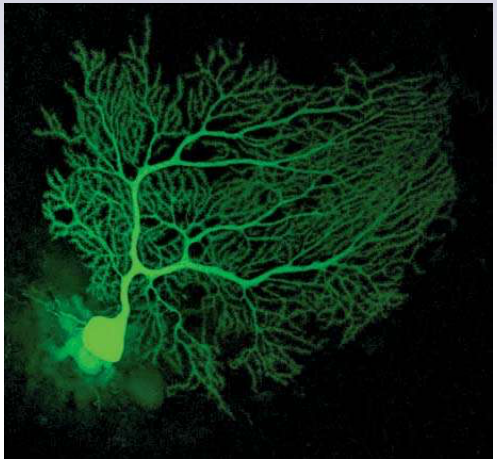




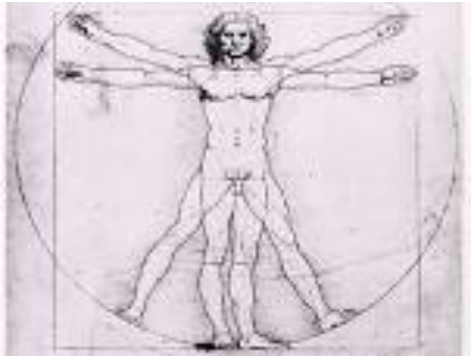
Mouse
(Maryann Martone
CCDB/NCMIR/UC San Diego)

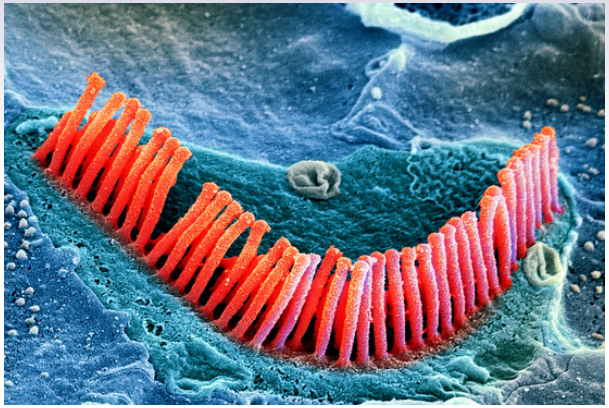
Pigeon
(drawing: Santiago Ramón y Cajal)

Different Genomic Code => Similar phenotype



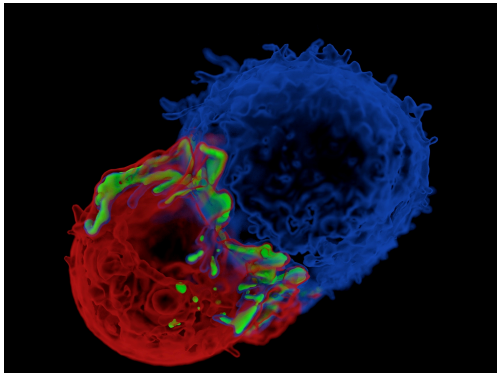
Purkinje cell



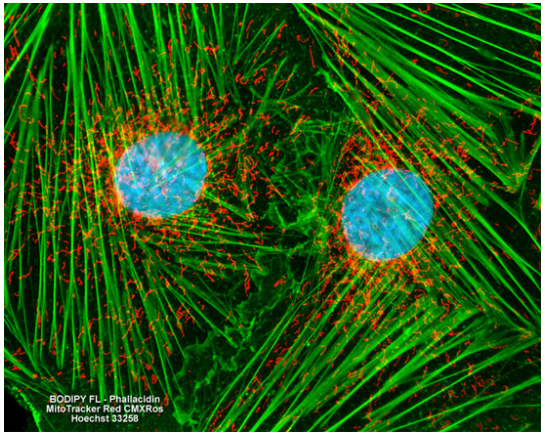


Hair cell

Same Genomic Code => Very Different Phenotype



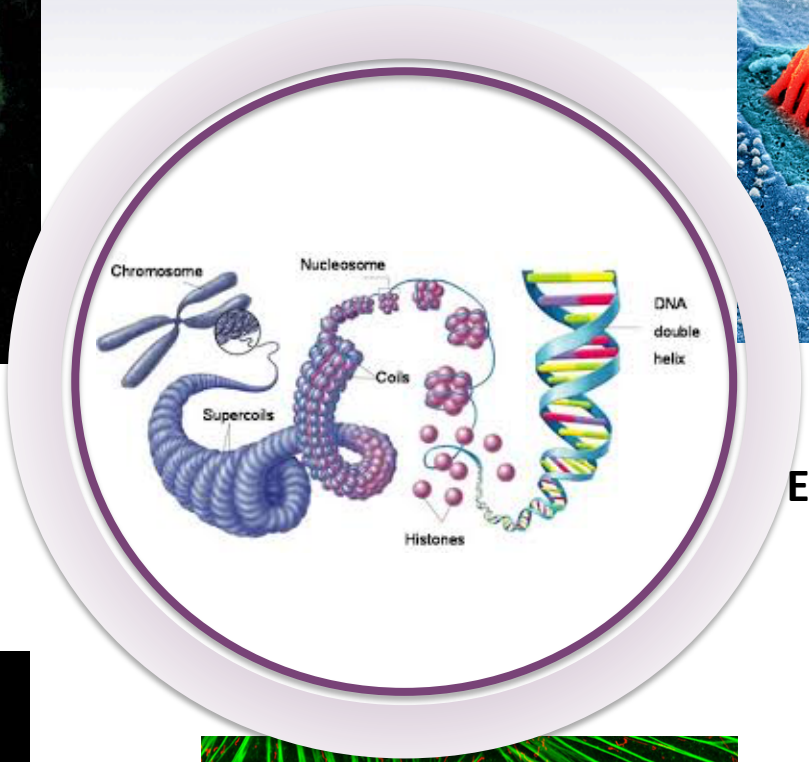
T cell (blue)



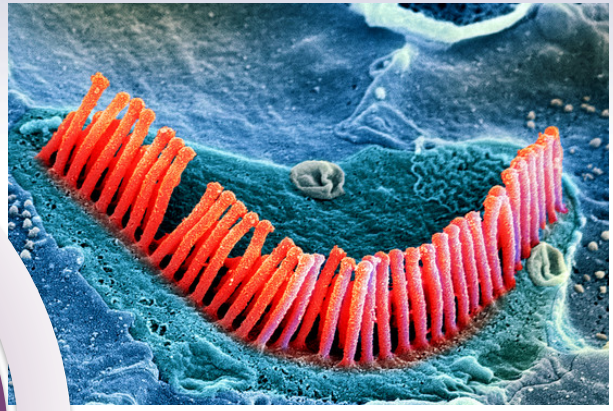
Smooth Muscle Fibroblast Cells

Ovum and sperms

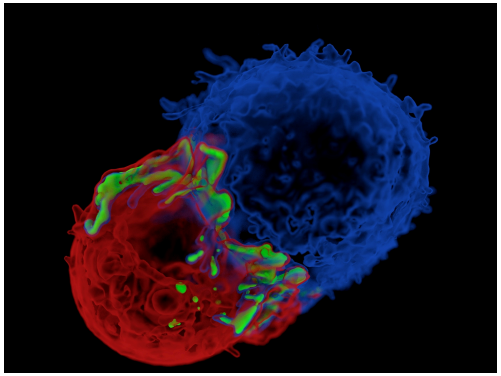
Purkinje cell



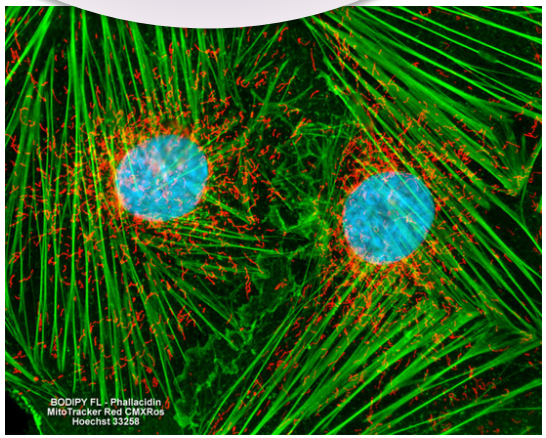
Epigenetic code



Hair cell

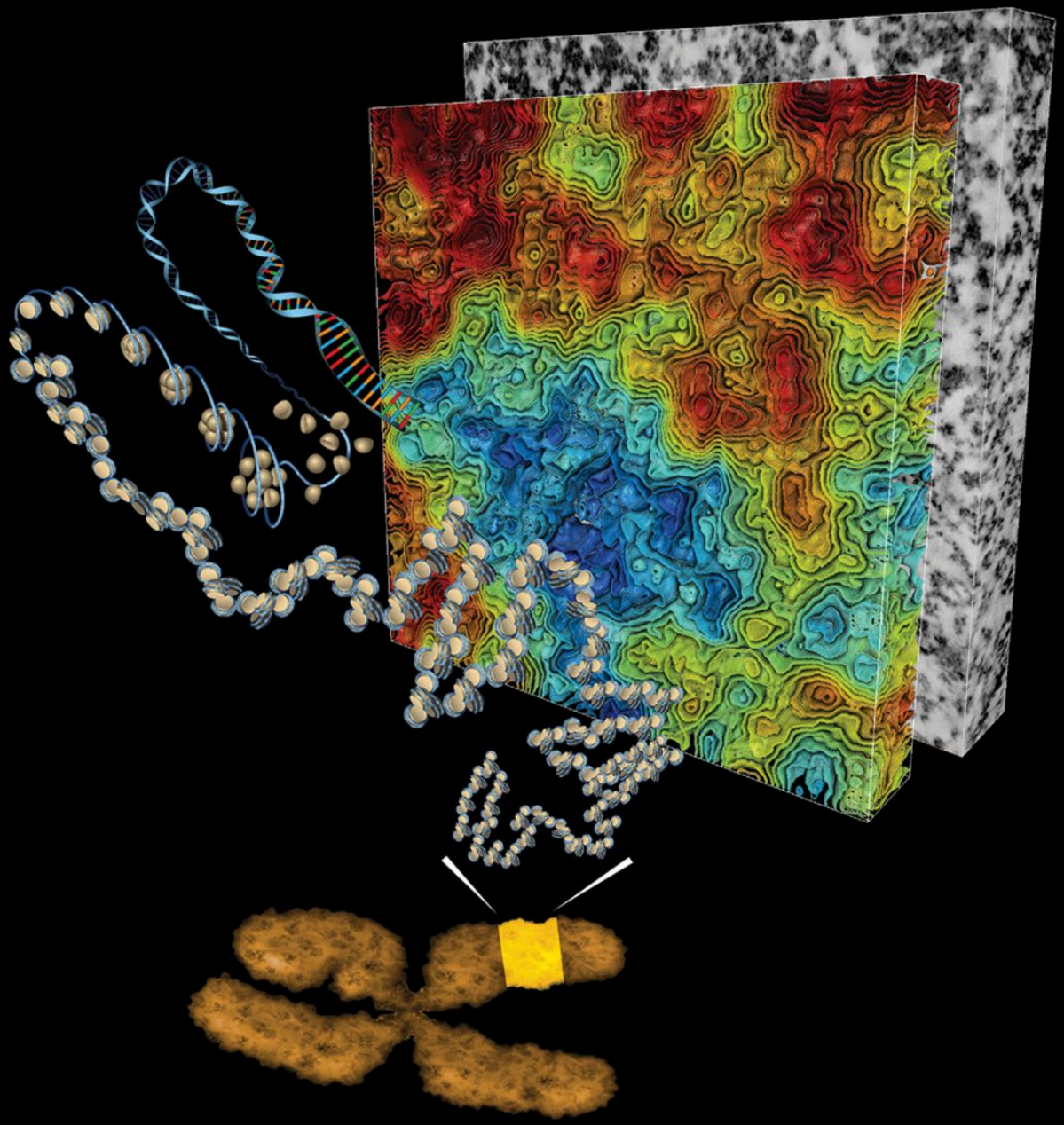


T cell (blue)

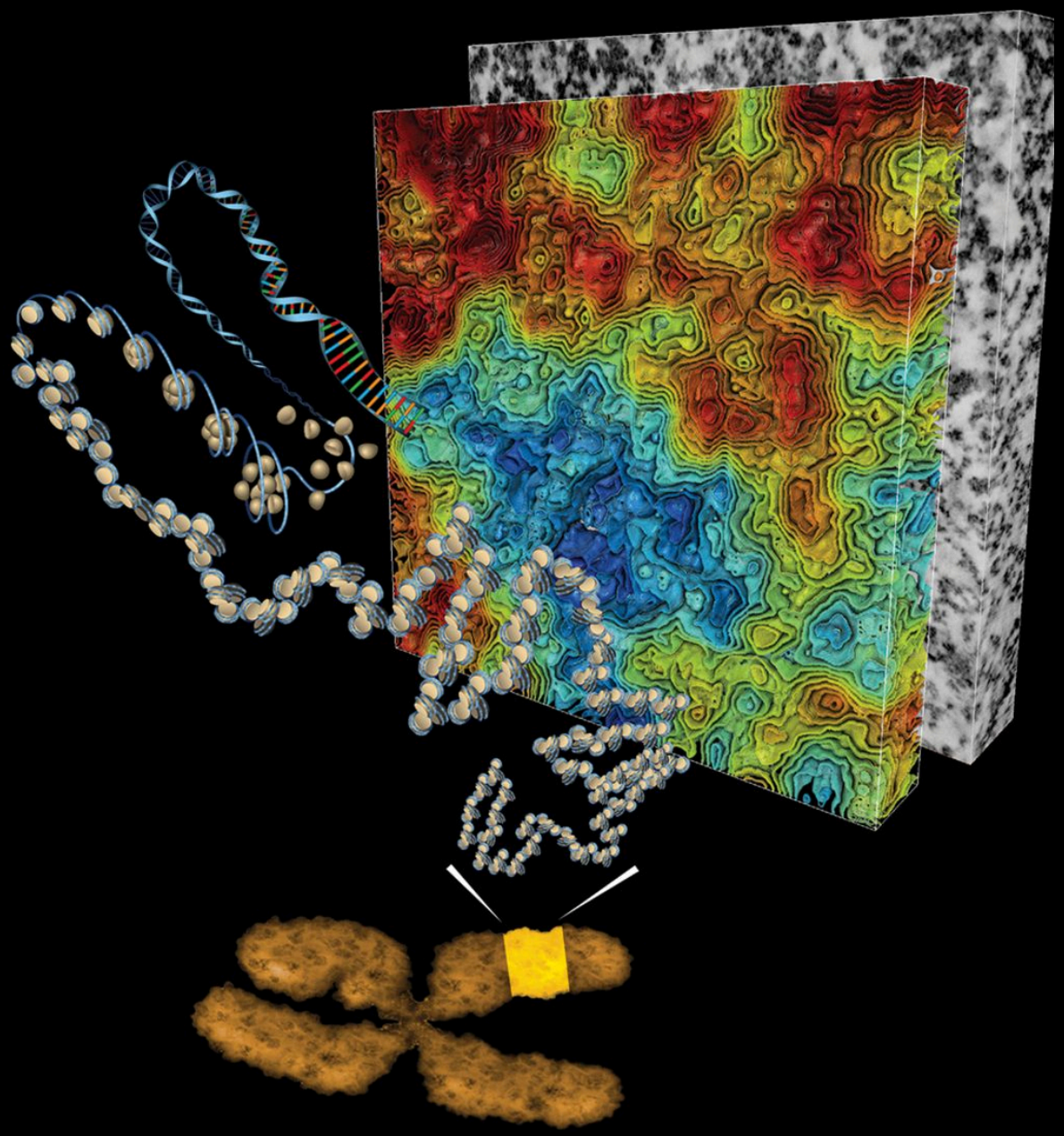


Smooth Muscle Fibroblast Cells

Ovum and sperms



Ou, H. D. *et al.* ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* (2017)



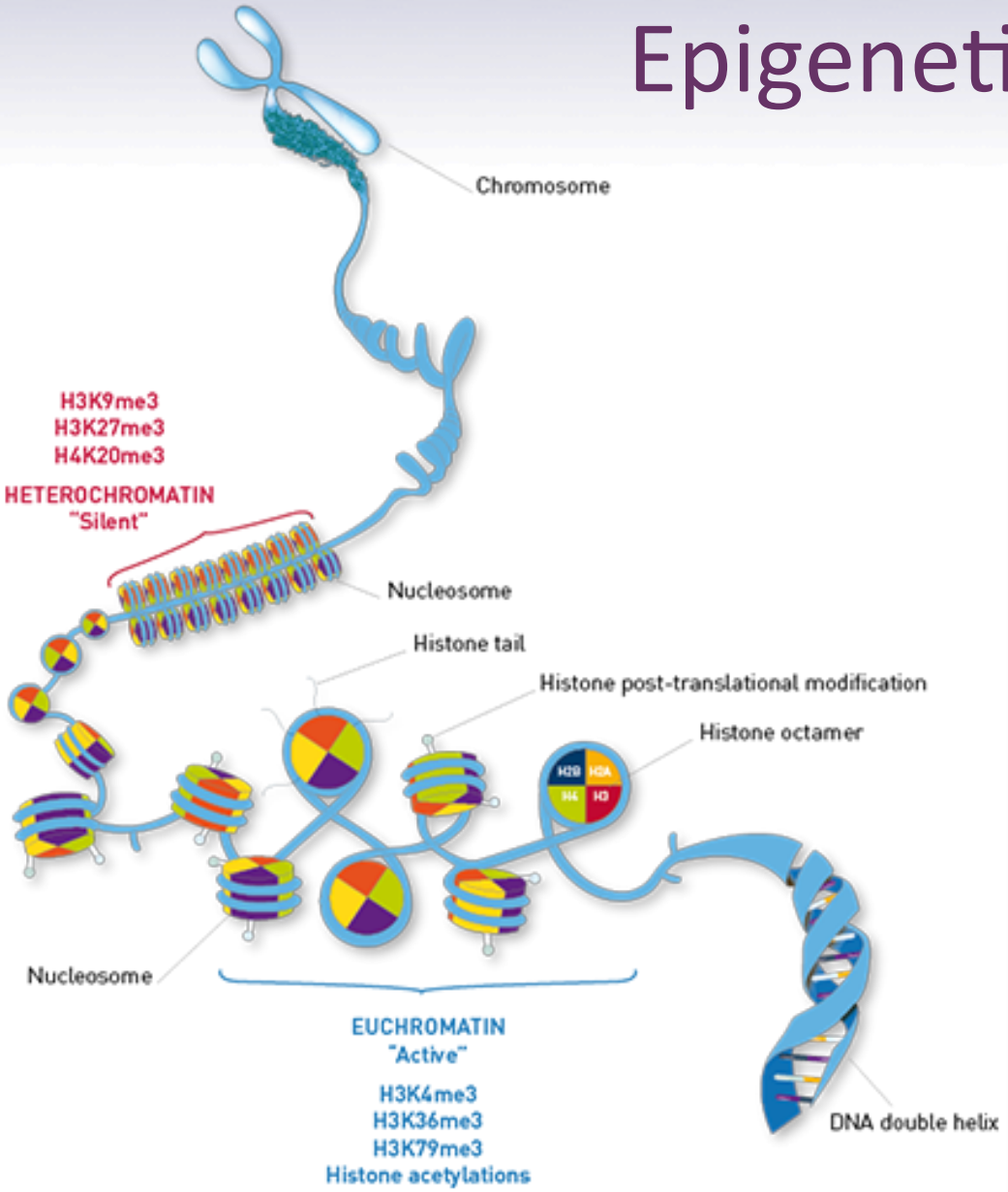
Epigenetics

R. Holliday (1990):
"… mechanisms that impart *temporal and spatial control* on the activities of all those gene required for the development of a complex organisms from zygote to the adult …"

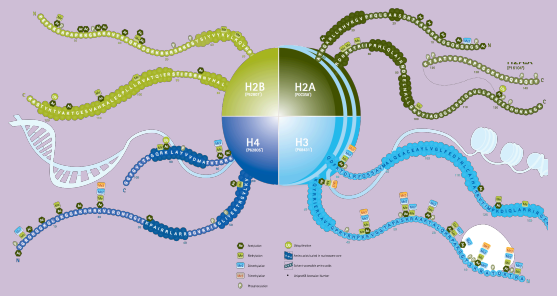
A.Riggs (1996):
"… mitotically and/or meiotically *heritable changes in gene function* that cannot be explained by changes in DNA sequence…."

A.Bird (2007):
"the structural adaptation of chromosomal regions so as to *register, signal or perpetuate* altered activity states "

Epigenetics

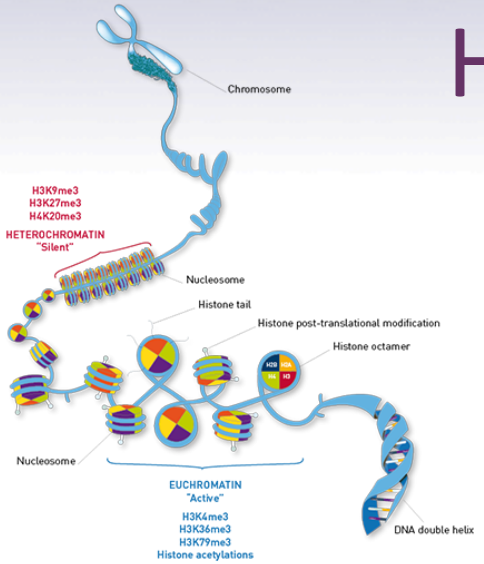


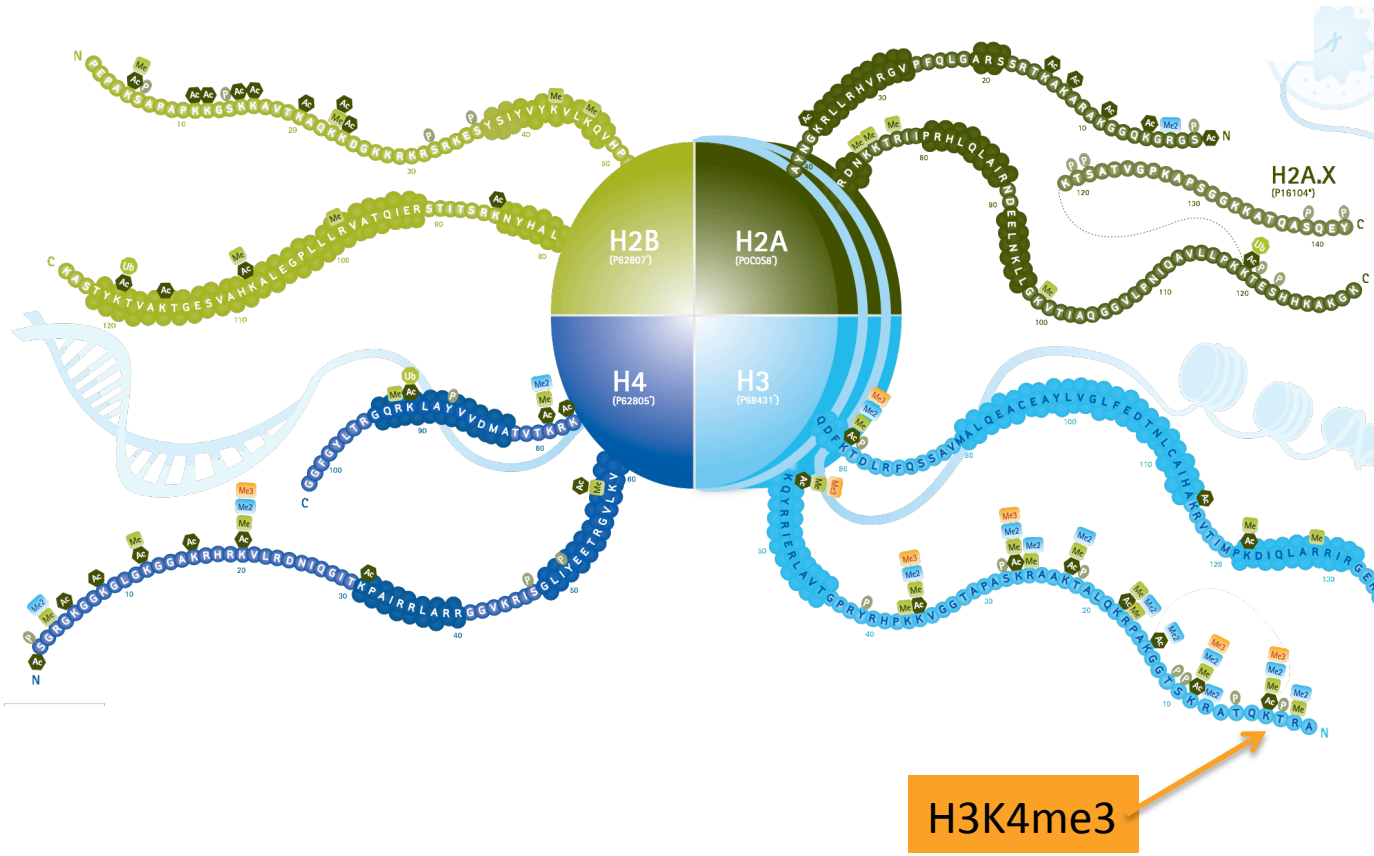
- DNA modifications (mC, hmC)
- ATP-dependent Chromatin remodeling
- **Histone Modifications**



https://www.diagenode.com/en/categories/histone-antibodies

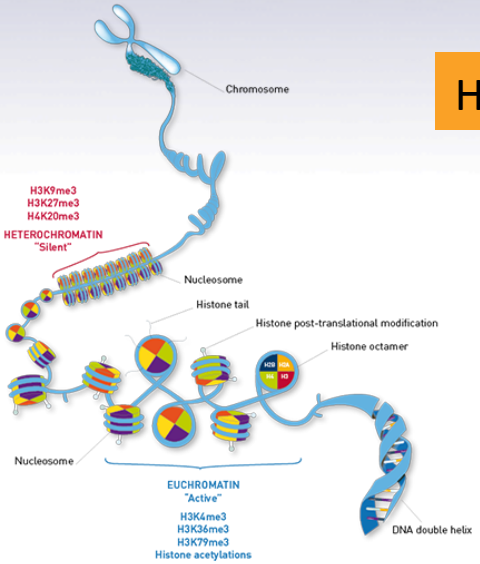
Histone Modifications



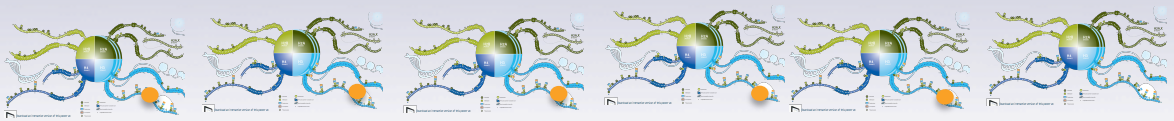


- Ac — Acetylation
- Me — Methylation
- Me2 — Dimethylation
- Me3 — Trimethylation
- P — Phosphorylation

http://www.merckmillipore.com/IE/en/life-science-research/genomic-analysis/Epigenetics-and-Nuclear-Function/Histone-Modification/73yb.qB.U9cAAAFOBtA1lTAF,nav?ReferrerURL=https%3A%2F%2Fwww.google.co.uk%2F&bd=1

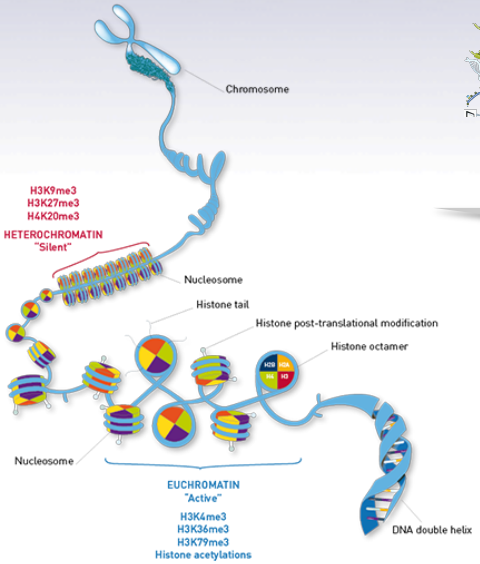


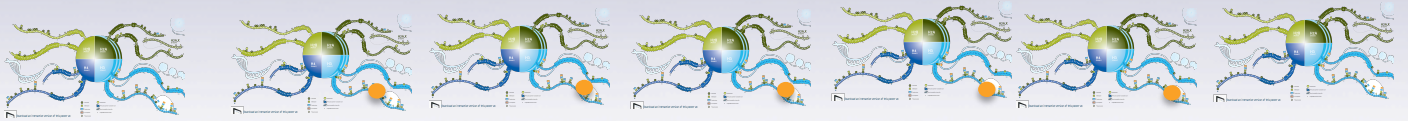
H3K4me3



ACAGTGAAGGATCGACAGTGAAGGATCGAAAGCTAGCCAGTAAGCTAGCCAGTACAGTGAAGCTAGCCAGT

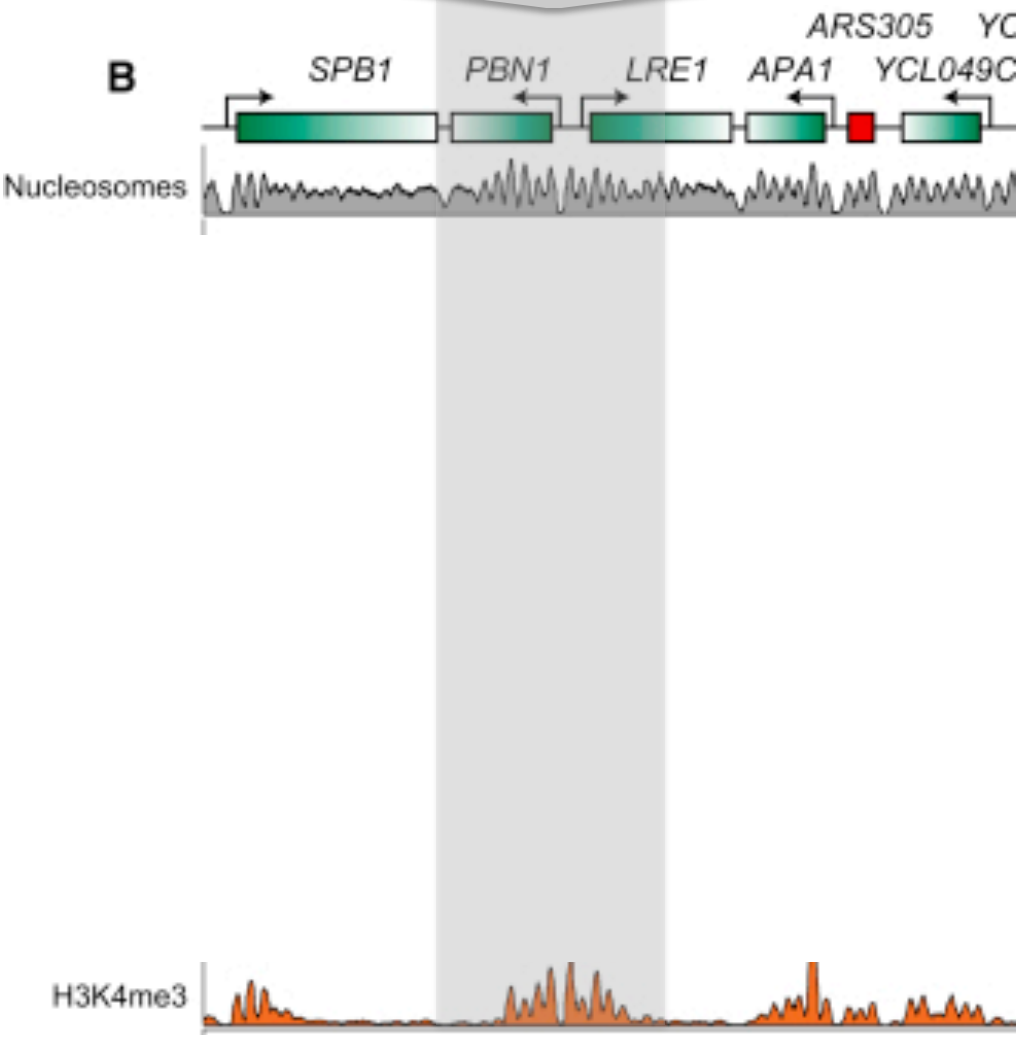
Genome





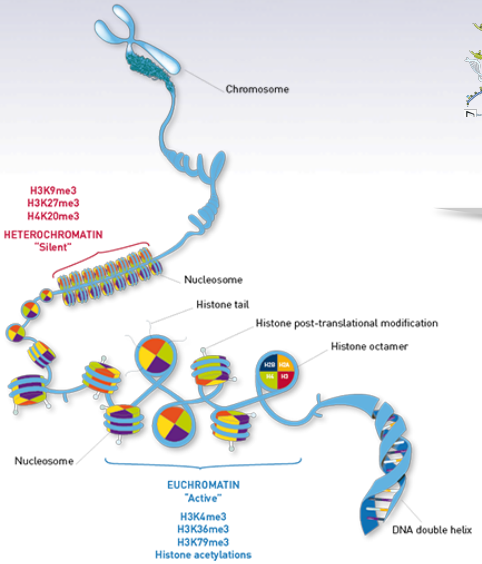
ACAGTGAAGGATCGACAGTGAAGGATCGAAAGCTAGCCAGTAAGCTAGCCAGTACAGTGAAGCTAGCCAGT

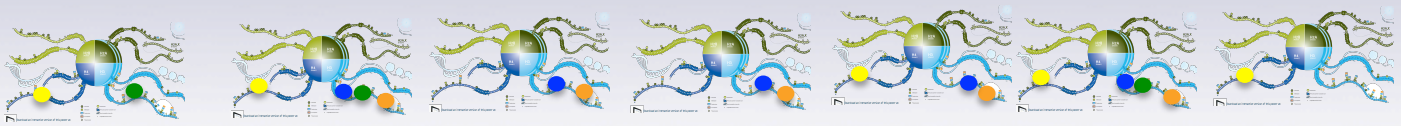
Genome



Weiner et al. Mol Cell, 2015

H3K4me3

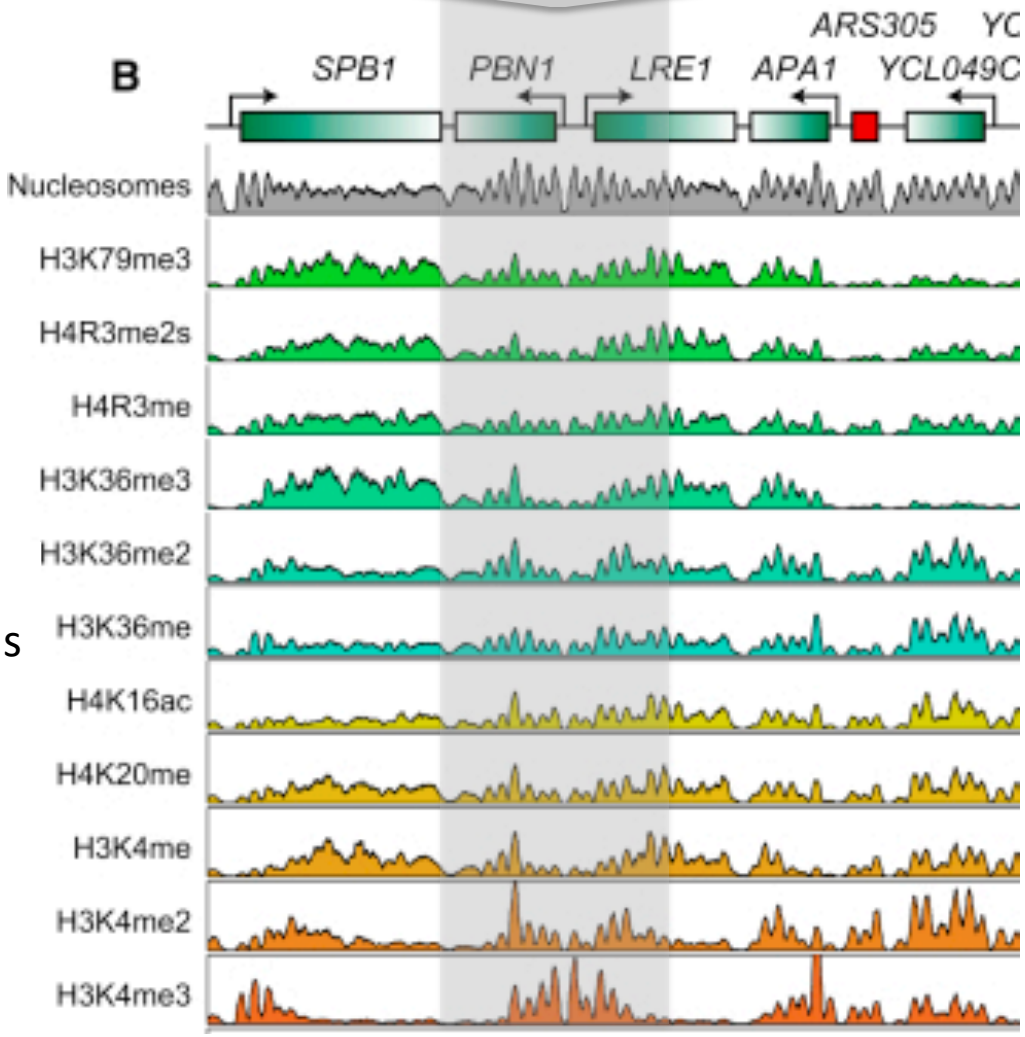




ACAGTGAAGGATCGACAGTGAAGGATCGAAAGCTAGCCAGTAAGCTAGCCAGTACAGTGAAGCTAGCCAGT

Genome

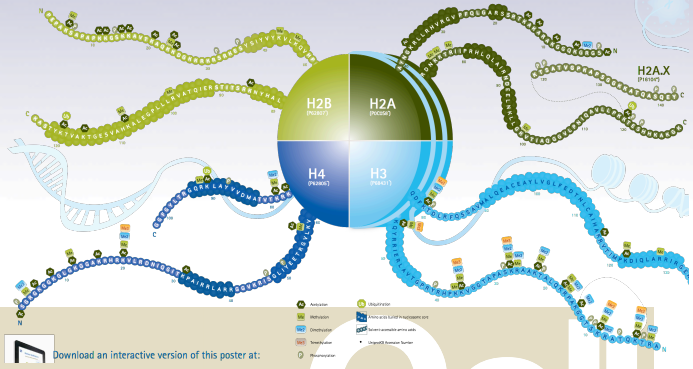
Histone modifications



Weiner et al. Mol Cell, 2015

H3K4me3

Histone Writers

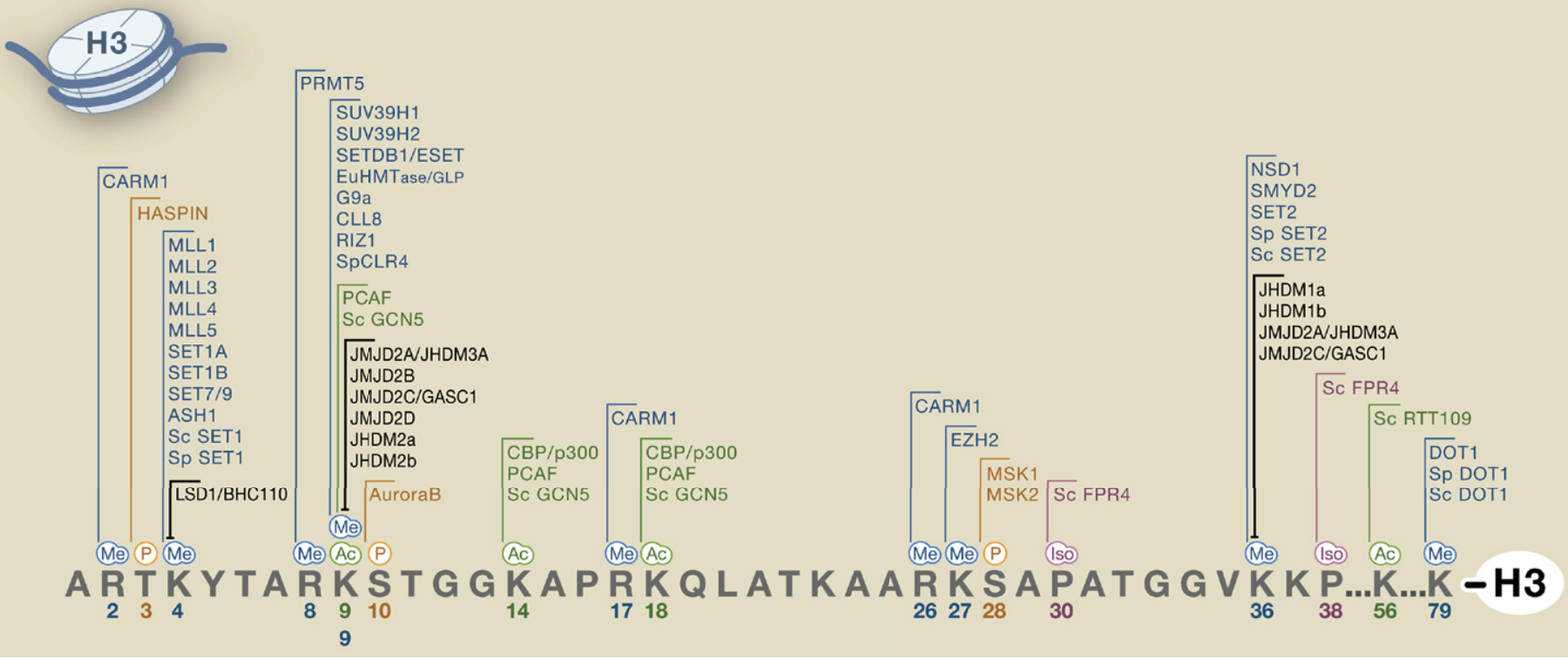


SnapShot: Histone-Modifying Enzymes

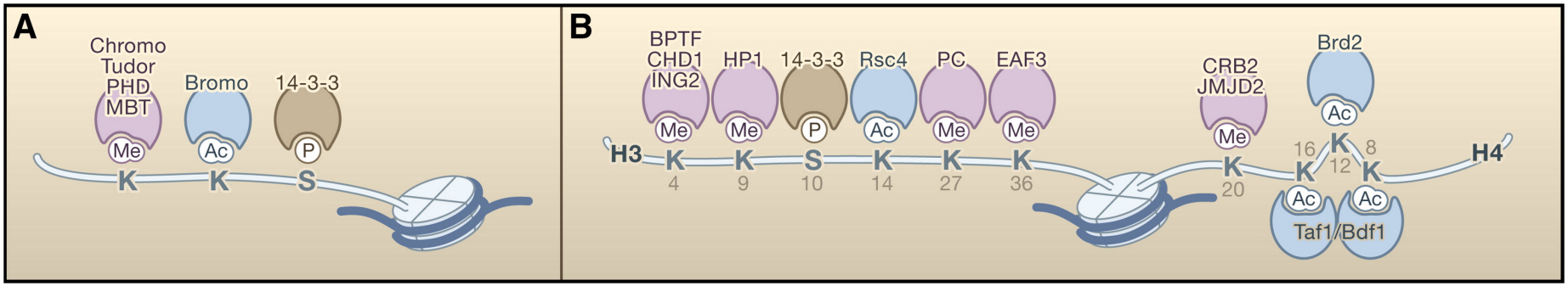
Tony Kouzarides
The Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK

Cell

Kouzarides, Cell. 2007

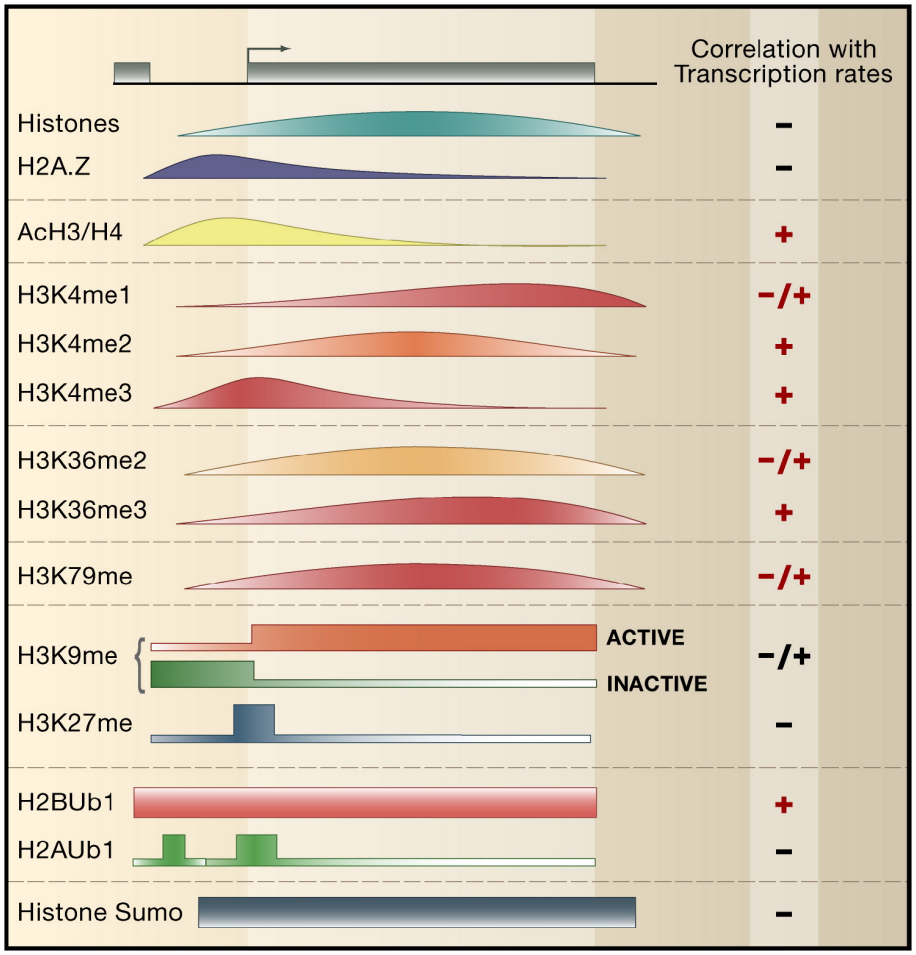


Readers of Histone Marks



Kouzarides, Chromatin Modifications and Their Function, Cell 2007

Histone Modifications Correlate with Transcription Activity

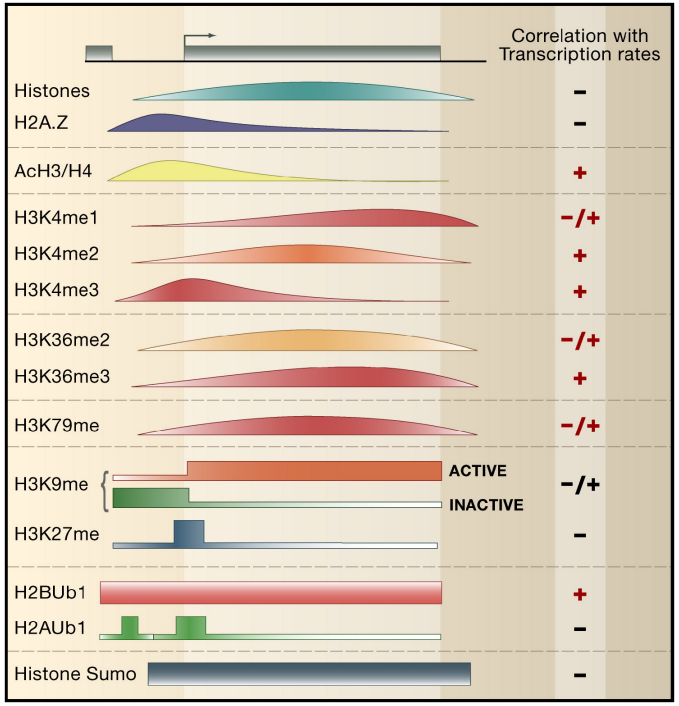


Li et al., The Role of Chromatin during Transcription, Cell,2007

Histone Code ?

Histone Modification Patterns

Expression / Transcription Output

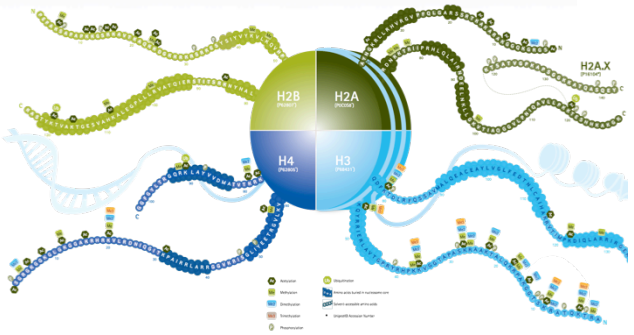


Histone Code ?

Complexity of Input :

- H3 contains 19 Lysines,
- can be mono-, di-, tri-methylated

4^19 = 280 billion different Lysine patterns
⇒ Huge “Alphabet”

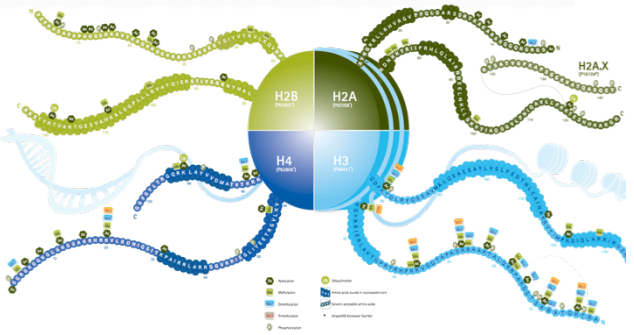


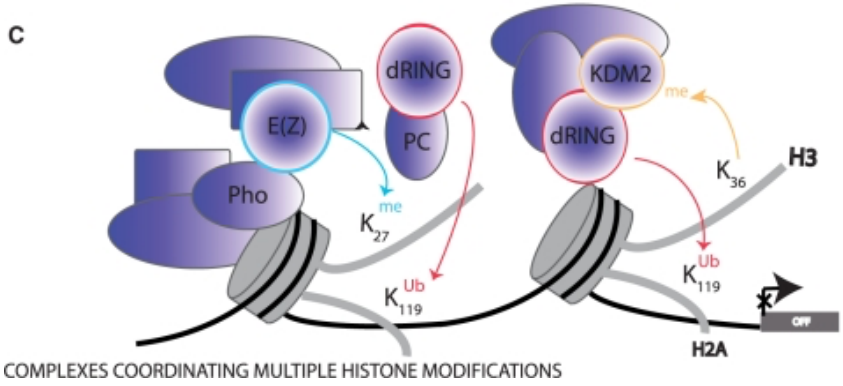
Histone Code ?

Complexity of Input :

- H3 contains 19 Lysines,
- can be mono-, di-, tri-methylated

4^19 = 280 billion different Lysine patterns
⇒ Huge “Alphabet”
Cross-talk between neighboring nucleosomes (potentially forming “words”)
⇒ Further increase in complexity ?





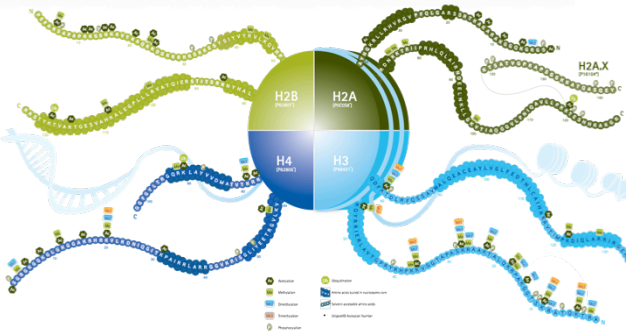
COMPLEXES COORDINATING MULTIPLE HISTONE MODIFICATIONS

Histone Code ?

Complexity of Input :

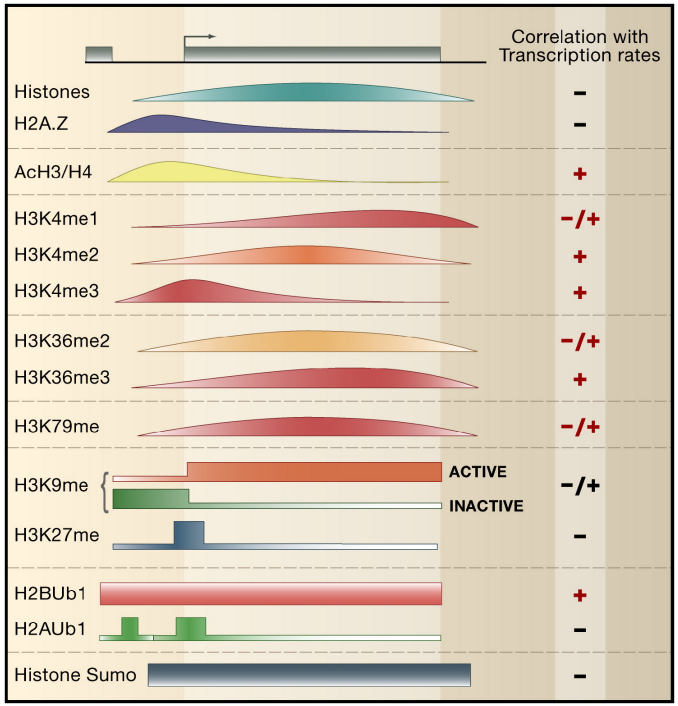
- H3 contains 19 Lysines,
- can be mono-, di-, tri-methylated

⇒ 4^19 = 280 billion different Lysine patterns



Complexity of Response:

- Heterochromatin vs Euchromatin
- Promoter vs Enhancer
- Activation vs Repression vs Bivalent



Talk Outline

Understanding the Complexity of Histone Modifications

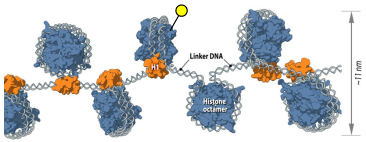
1. **ChIP - Seq Analysis**
2. Detecting Differential Modified Regions
3. Identifying Chromatin States
4. Data Integration

Future Directions

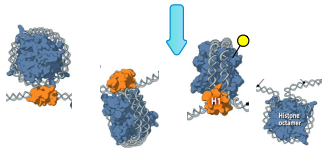
1. From Correlation to Causality
2. Single-Cell ChIP-Seq
3. Time Course Data

ChIP-Seq

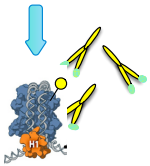
DNA



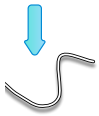
cross-linking



DNA fragmentation



Enrichment with specific antibody (ChIP-)



sequencing (-seq)

alignment

DNA

ACGCGATCAGGGCGACGCGATCAGAGCTAACGGTGCAACGGGTAC

ChIP-Seq

to determine

- position of histone modifications

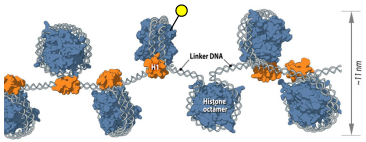
- stability of mark

mark strength
N = 9 counts

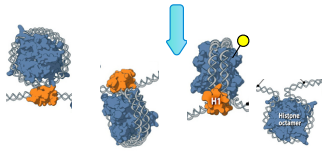
enriched region:
'peak'

ChIP-Seq

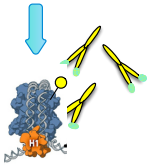
DNA



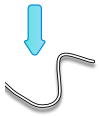
cross-linking



DNA fragmentation



Enrichment with specific antibody (ChIP-)



sequencing (-seq)

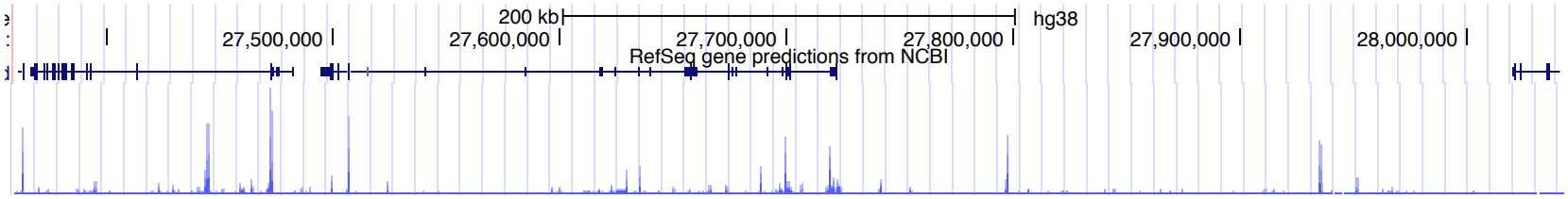
alignment

ChIP-Seq

to determine

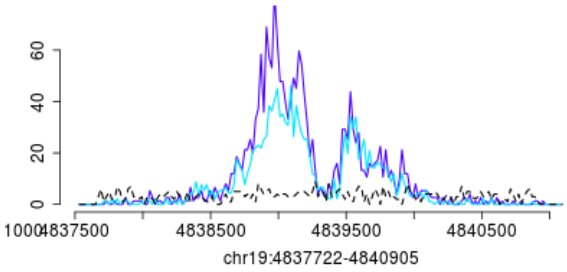
- position of histone modifications

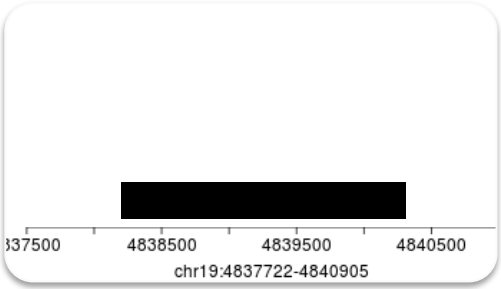
- stability of mark



ChIP-Seq Computational Pipeline

ChIP-Seq Analysis



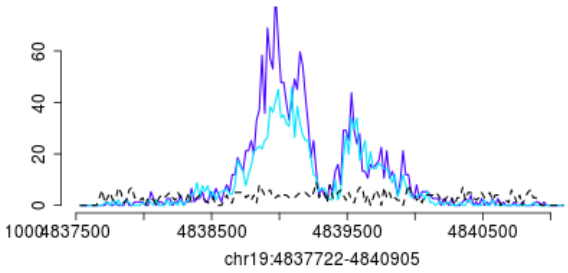


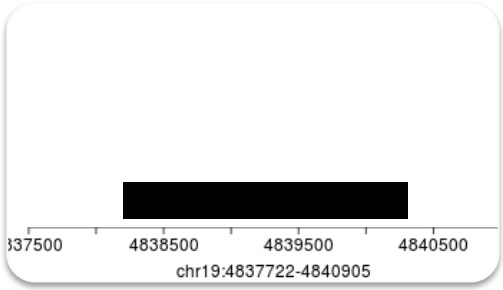
Detecting Enriched Regions

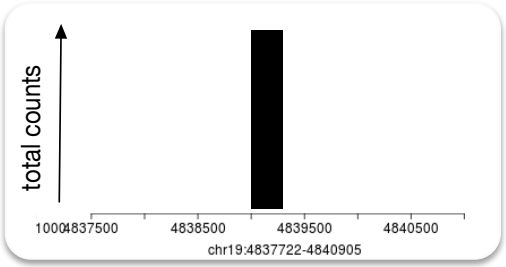
Presence / Absence

(Binary signal : 1/0)

ChIP-Seq Analysis







Detecting Enriched Regions

Presence / Absence

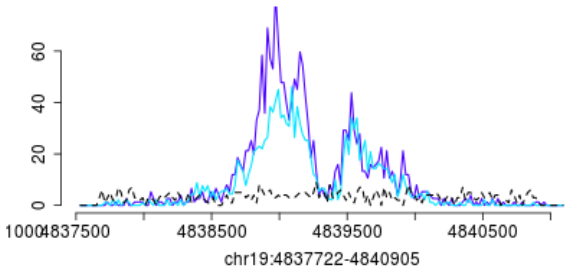
(Binary signal : 1/0)

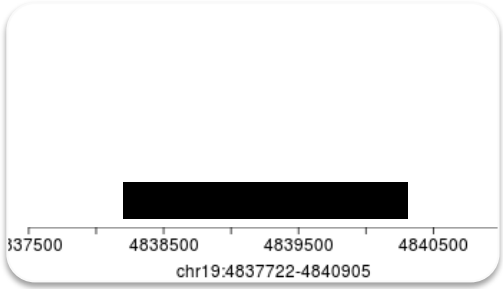
Quantifying Enrichments

Sum of Counts

(Count Signal: N)

ChIP-Seq Analysis

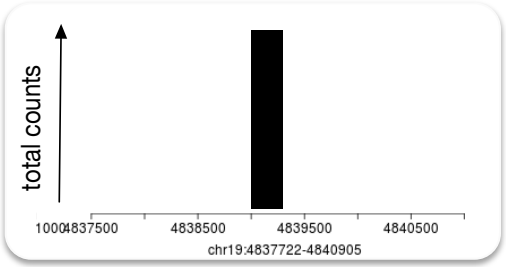




Detecting Enriched Regions

Presence / Absence

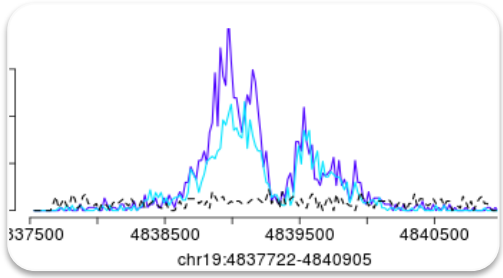
(Binary signal : 1/0)



Quantifying Enrichments

Sum of Counts

(Count Signal: N)

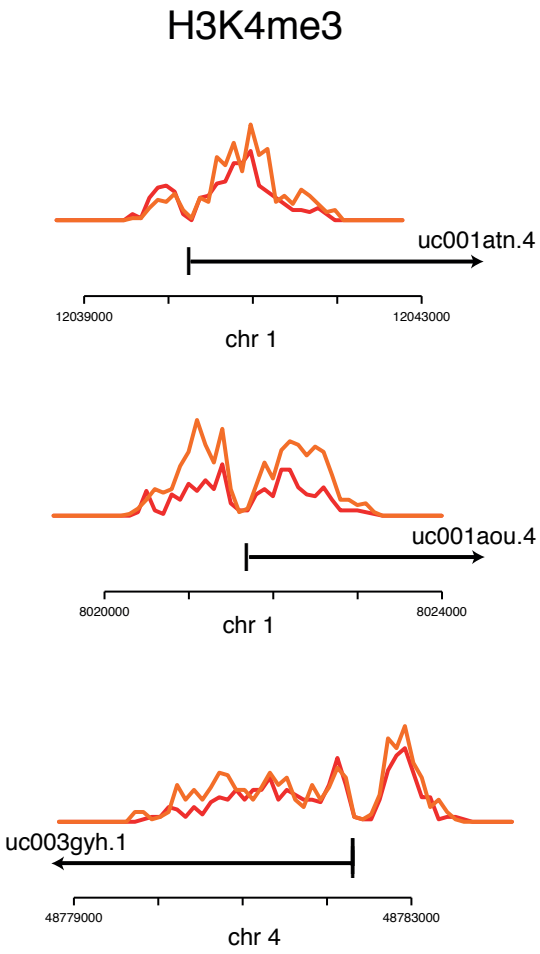


Shape Analysis

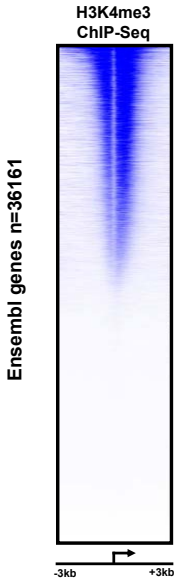
Distribution of Reads

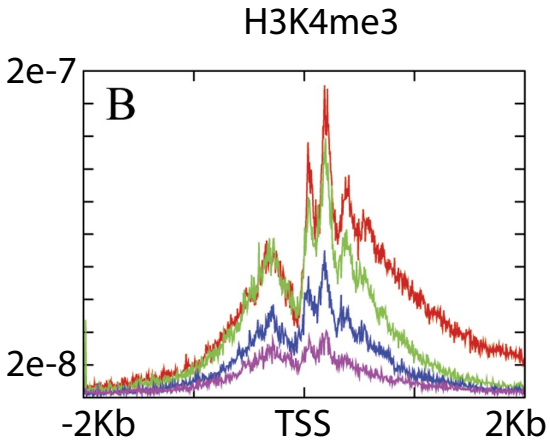
(Complex Signal: N^L)

Peak Alignment and Clustering



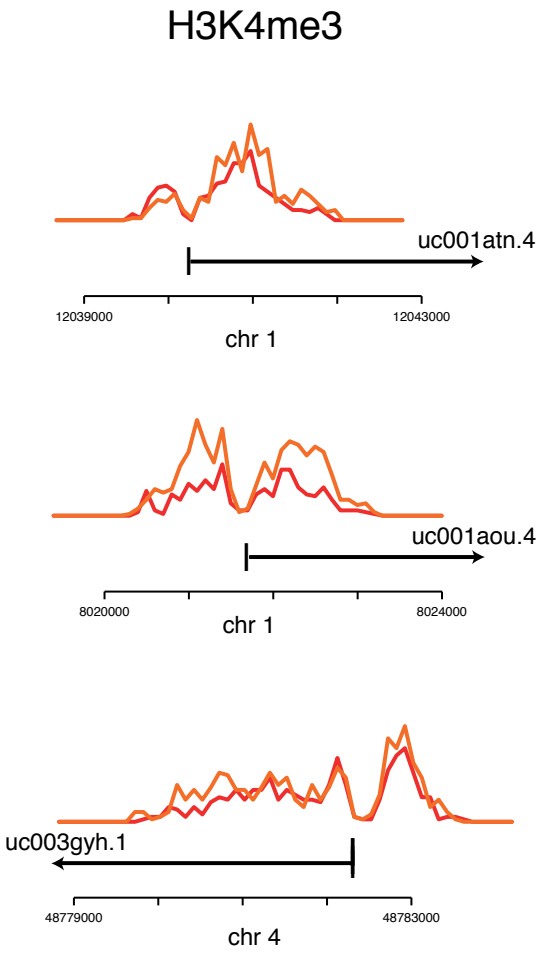
- *shape* of epigenomic patterns mark functional features (promoters, enhancers)





Barski, Cell 2007

Peak Alignment and Clustering

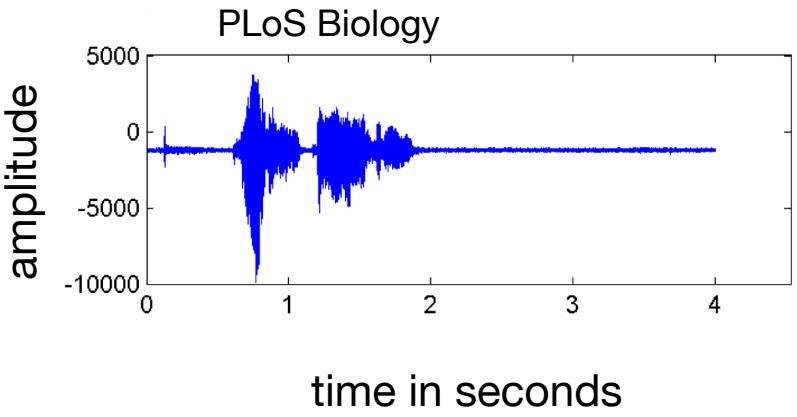


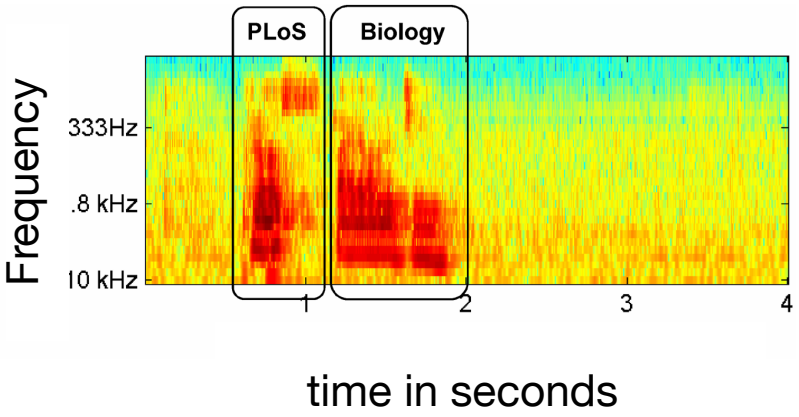
- *shape* of epigenomic patterns mark functional features (promoters, enhancers)
- epigenomic marks have local variation, which may be irrelevant for their function
- Can with classify these three peaks as the same pattern?

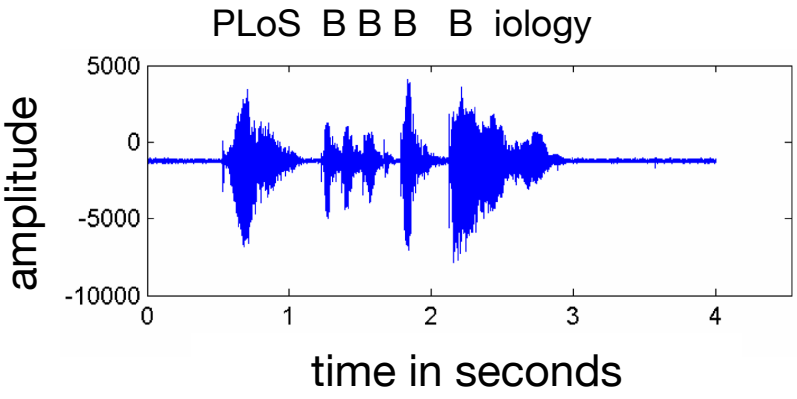
=> Dynamic Genome Warping (DGW)

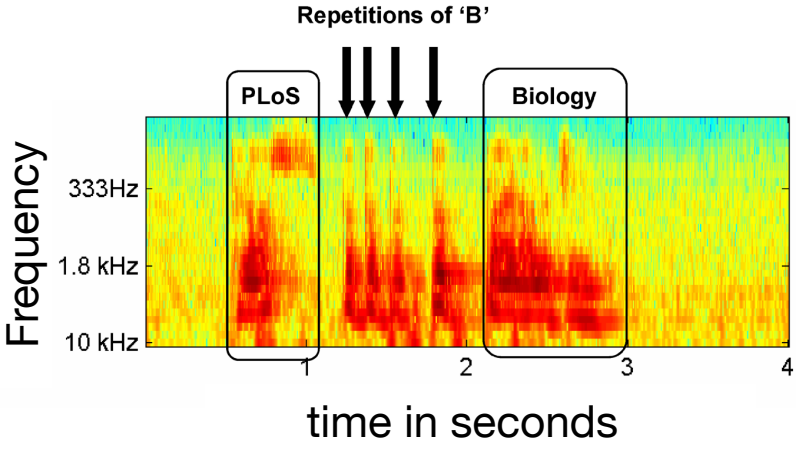
Analogy: speech recognition

Dynamic Time Warping (DTW)





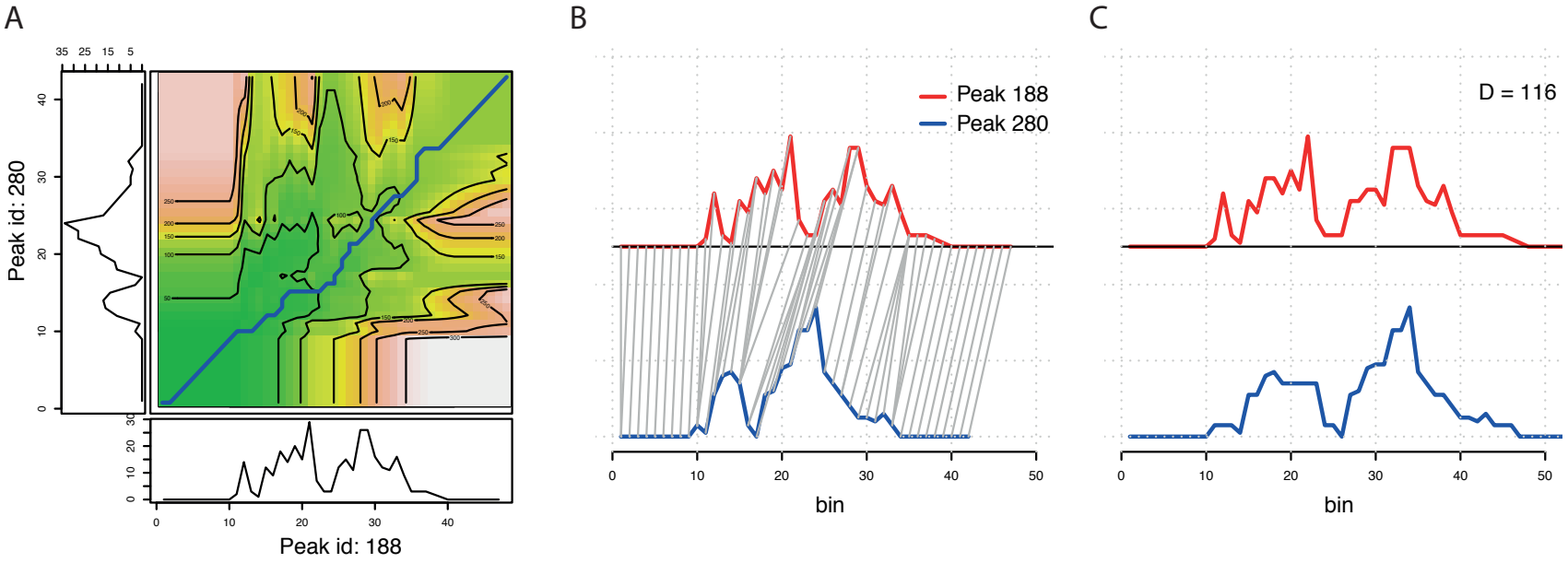




Büchel et al. , PLoS Biology 2004

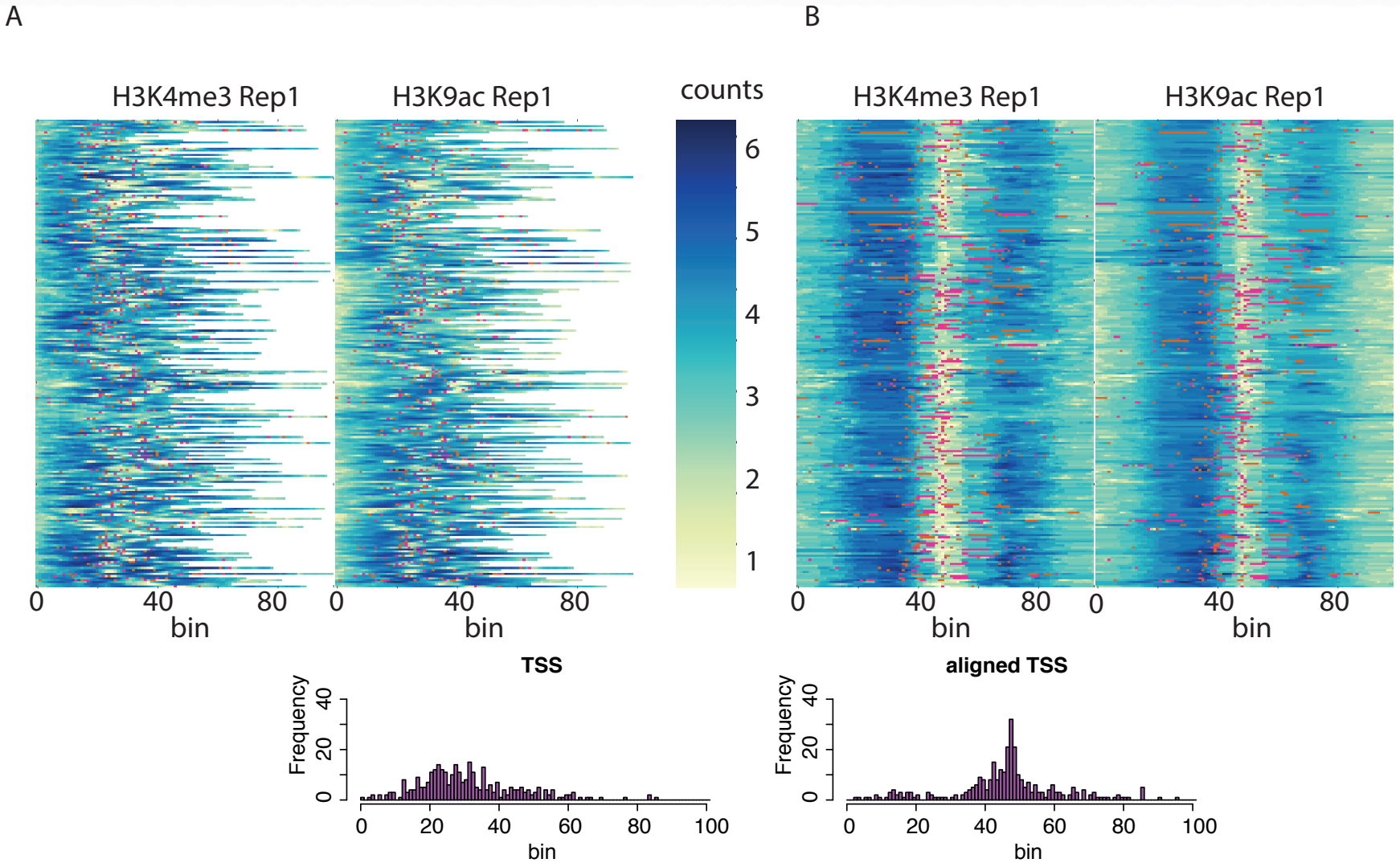
Peak Alignment Using DGW

=> Dynamic Genome Warping (DGW)



Lukauskas, et al. 2016

Results: DGW aligns genomic landmarks



Lukauskas, et al. 2016

Talk Outline

Understanding the Complexity of Histone Modifications

1. ChIP - Seq Analysis
2. **Detecting Differential Modified Regions**
3. Data Integration
4. Identifying Chromatin States

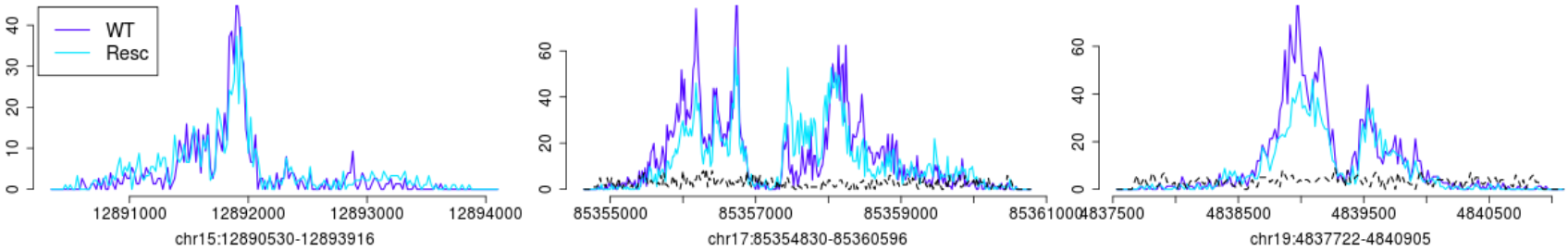
Future Directions

1. From Correlation to Causality
2. Single-Cell ChIP-Seq
3. Time Course Data

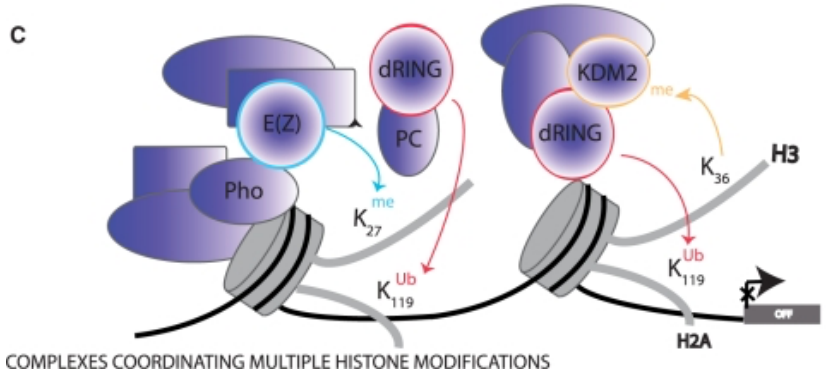
Detecting Differential Modified Regions

- Comparing sets of enriched regions
  - Presence/absence of mark (DiffBind)
- Comparing average levels of modification (Diffbind)
- Comparing Shape of Modification (MMDiff)

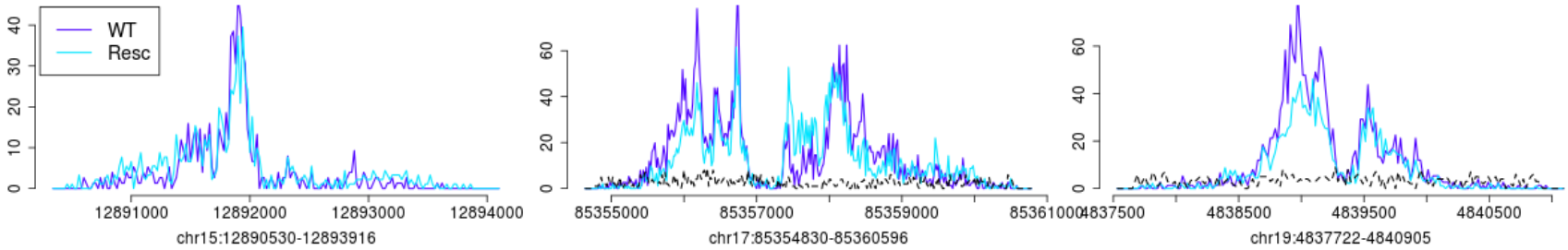
H3K4me3 at selected promoters in mES cells



Sub-structure of binding peaks are remarkably conserved between experiments.



H3K4me3 at selected promoters in mES cells



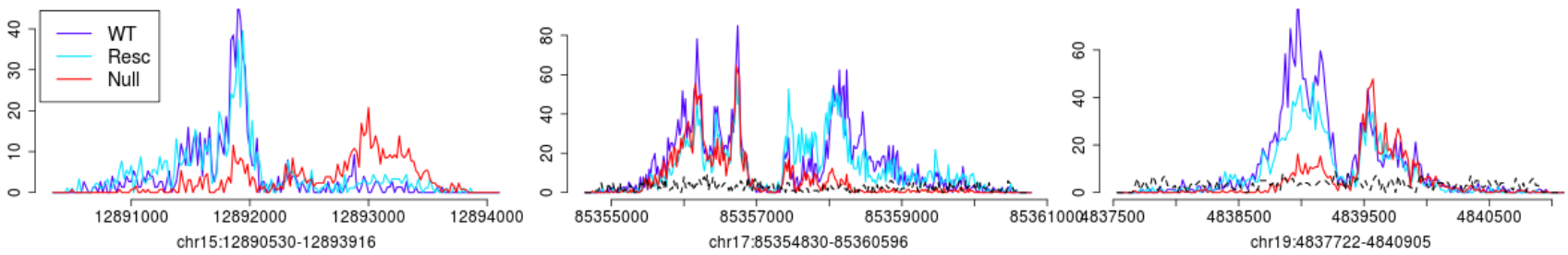
Sub-structure of binding peaks are remarkably conserved between experiments.

Is there biological function encoded in the shape of the peak?

How are they established?

=> Loss of function experiment.

H3K4me3 loss in a Cfp1 mutant



The mark is not lost in a homogenous way.
Some parts are not affected.

Computational Challenge:
Detect Differences in Shapes of Peaks rather than intensity

Standard approach:
extract a summary statistic (total counts)
Univariate test (e.g. negative binomial)

⇒ Low power

Our Idea:
Sequencing itself is a form of *sampling* an *unknown distribution on the genome*
Number of drawn samples is identical to the number of reads observed in a peak

=> Greatly increased power

Re-formulate the test question

Suppose for a peak we are given

- n observations (i.e. reads) in data set *s* (disease)

  $X^s = x_1^s, \ldots, x_n^s$

- m observations in data set *s'* (control)

  $X^{s'} = x_1^{s'}, \ldots, x_m^{s'}$

where x^s, $x^{s'}$ random variables drawn *i.i.d.* from **unknown** probability distributions *p* and *p'*

Can we decide whether $p = p'$?

MMDiff

- MMD: **Maximum Mean Discrepancy**
- Kernel-based non-parametric test (Gretton et al., 2008, 2012)
- retains higher order information within the testing procedure

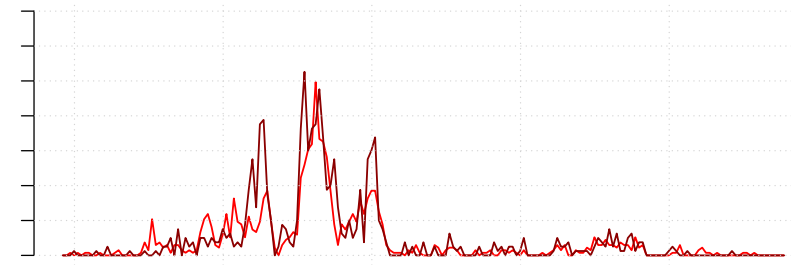
concept

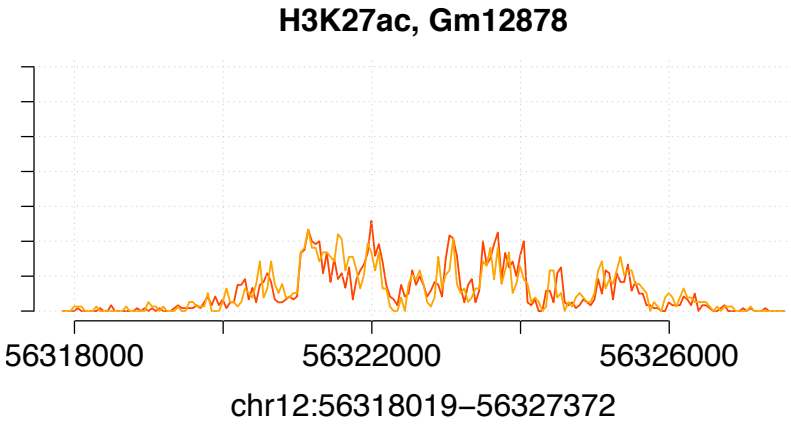
Define feature map, which maps the distributions into a high dimensional reproducing Kernel Hilbert Space (RKHS)

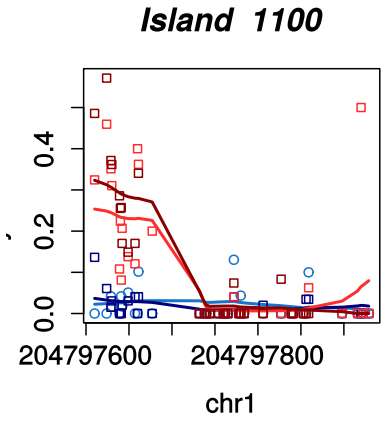
In this space, two distributions are identical if and only if their kernel means are identical

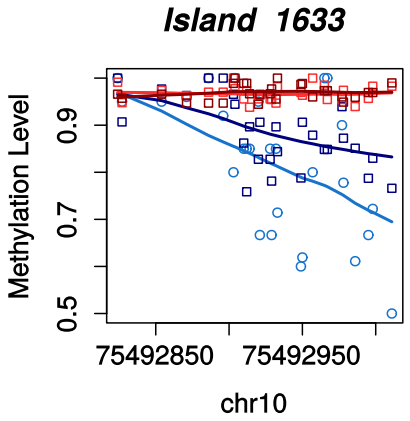
Distance between means is a good quantitative measure for difference between two distributions

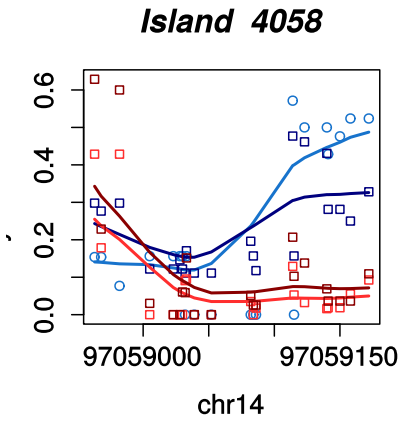
Results

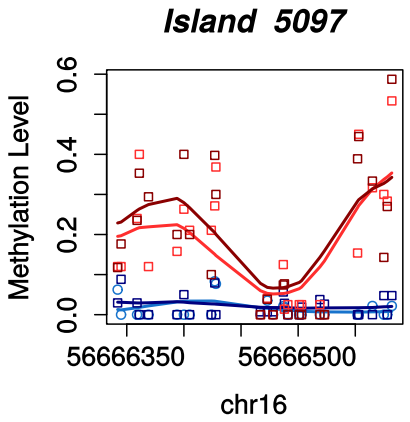












Schweikert et al., 2013

Mayo et al., 2015

Talk Outline

Understanding the Complexity of Histone Modifications

1. ChIP - Seq Analysis
2. Detecting Differential Modified Regions
3. **Data Integration**
4. Identifying Chromatin States

Future Directions

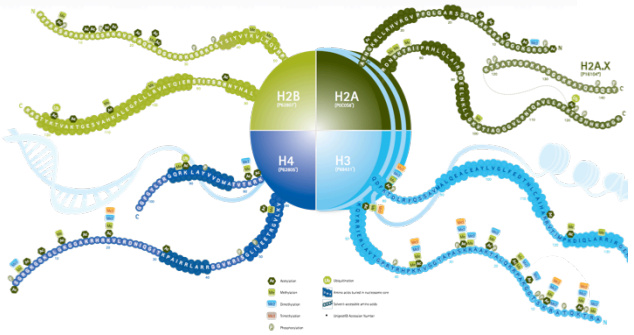
1. From Correlation to Causality
2. Single-Cell ChIP-Seq
3. Time Course Data

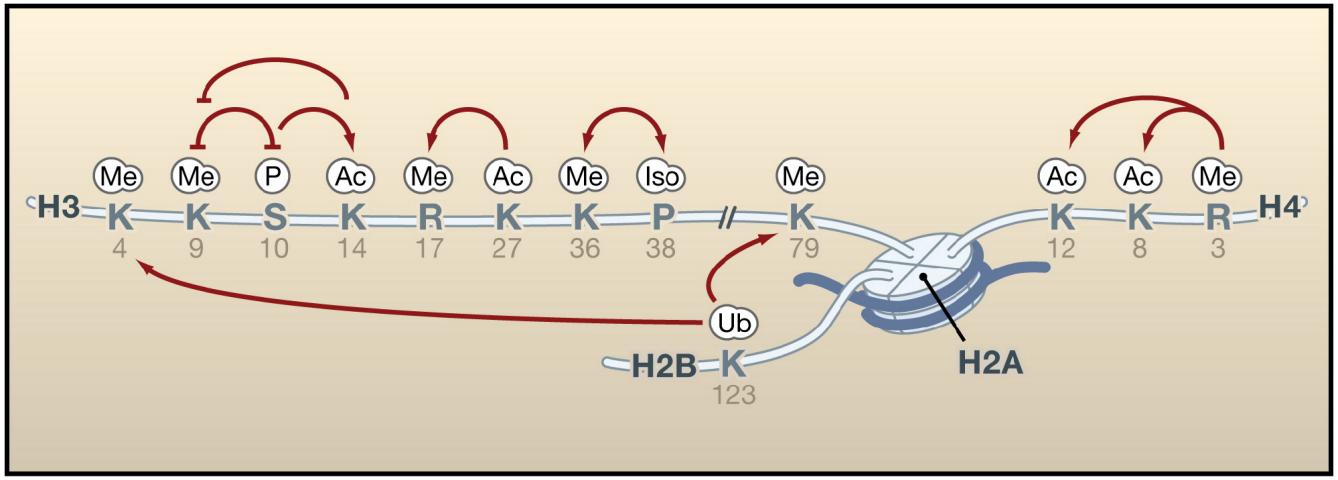
Cross-talk between different histone modifications

Complexity of Input :

- H3 contains 19 Lysines,
- can be mono-, di-, tri-methylated

⇒ 4^19 = 280 billion different Lysine patterns



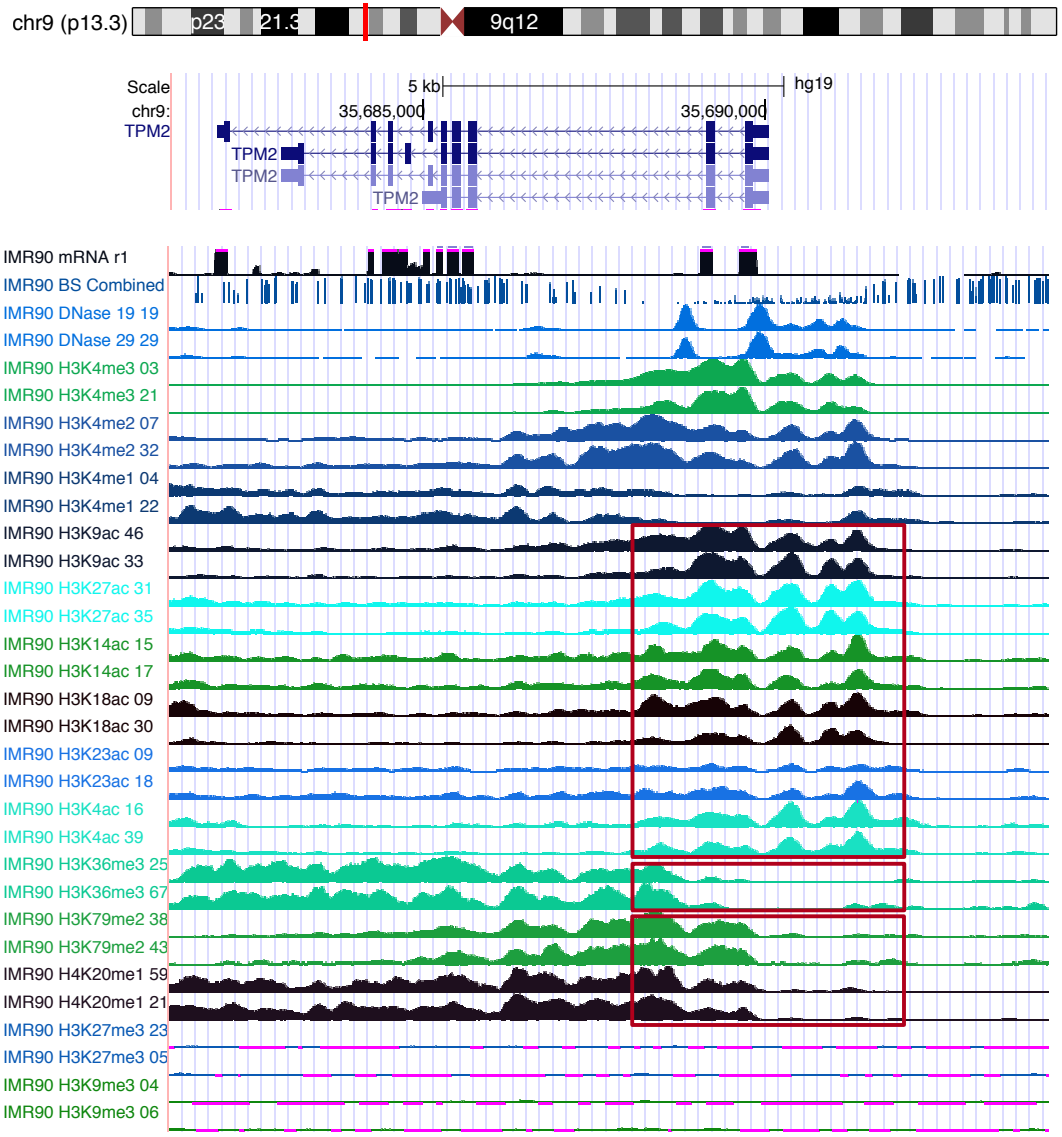


Kouzarides, Chromatin Modifications and Their Function, Cell 2007

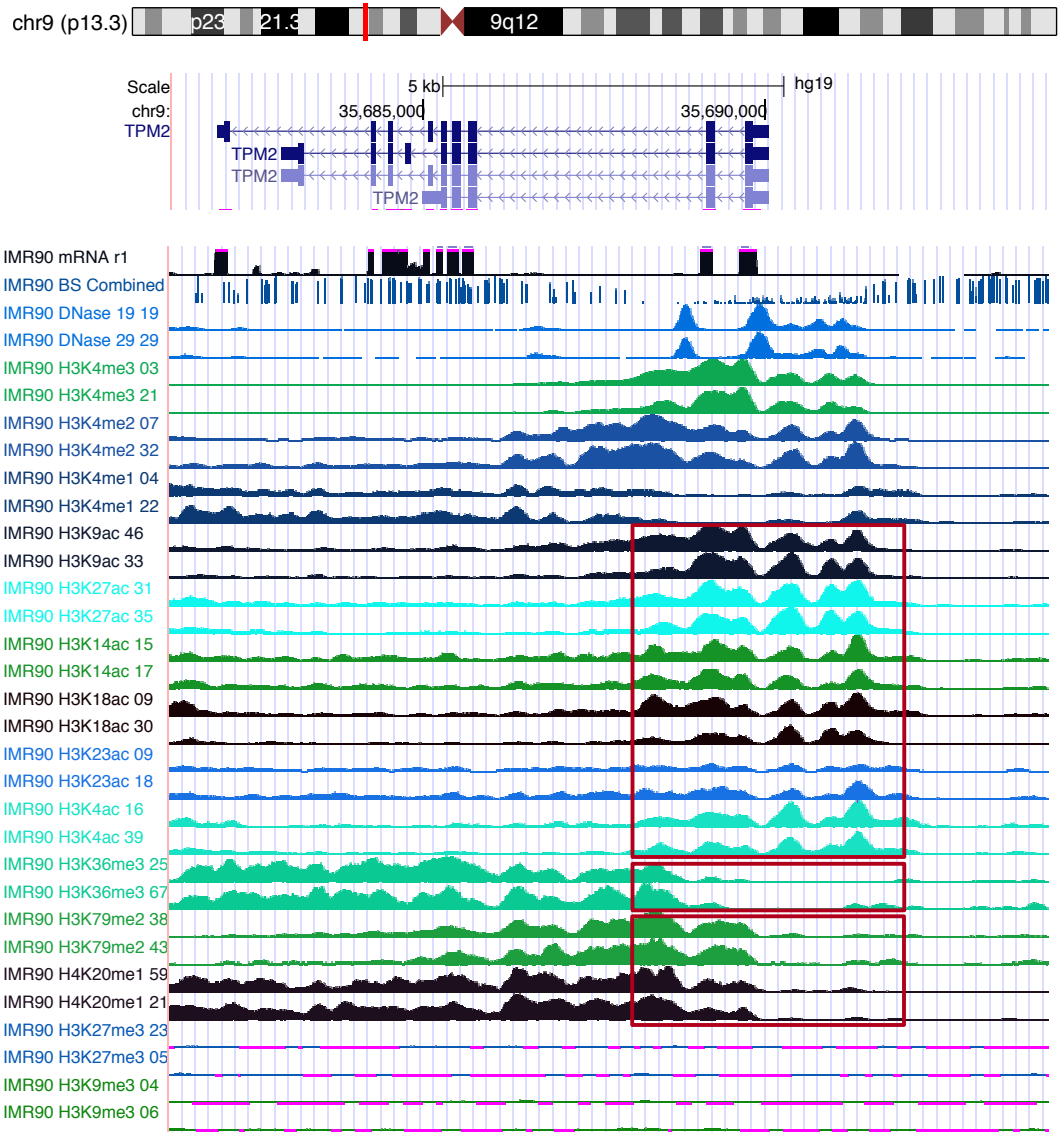
⇒ Individual marks are not independent

⇒ Reduction in Complexity

Epigenomic Crosstalk

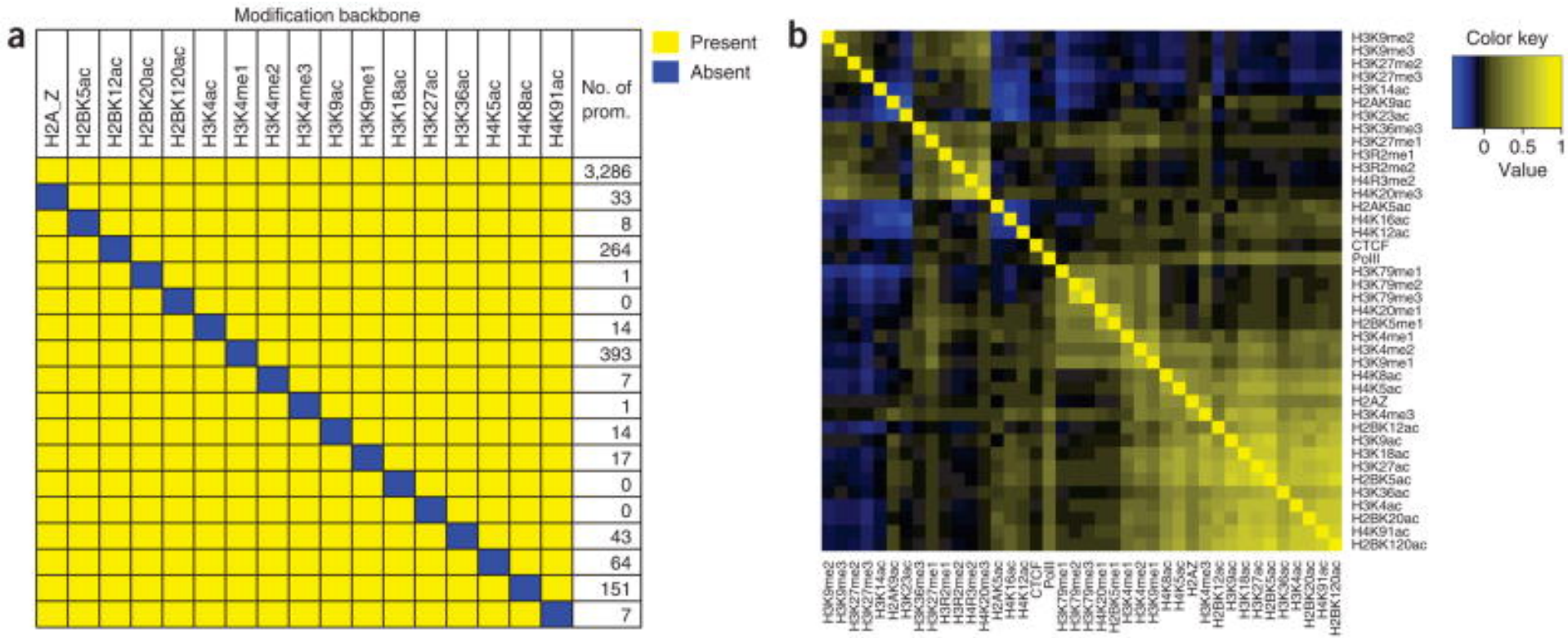


Epigenomic Crosstalk



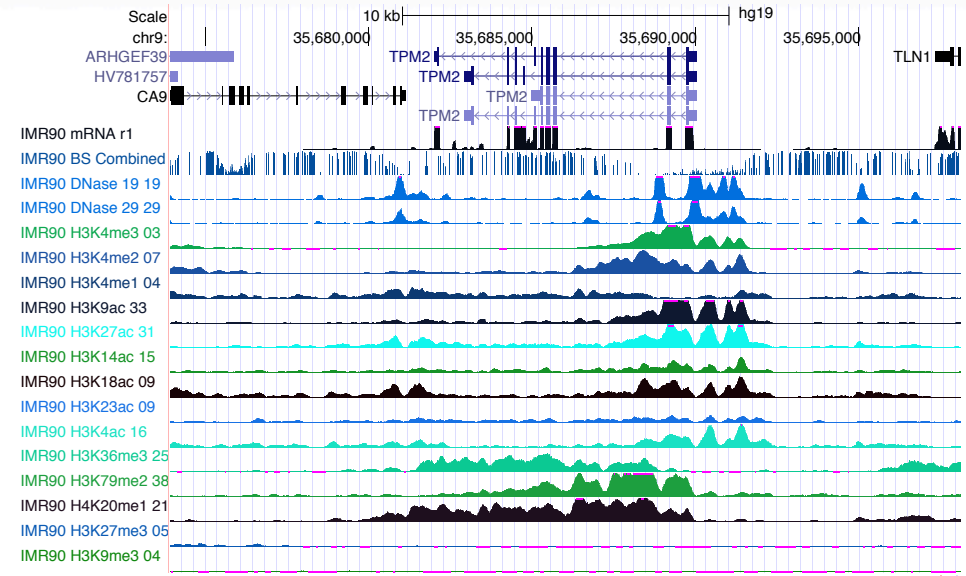
counts	Binary: presence/absence
145	1
135	1
123	1
132	1
145	1
143	1
23	0
60	0
54	0
53	0

Epigenomic Crosstalk



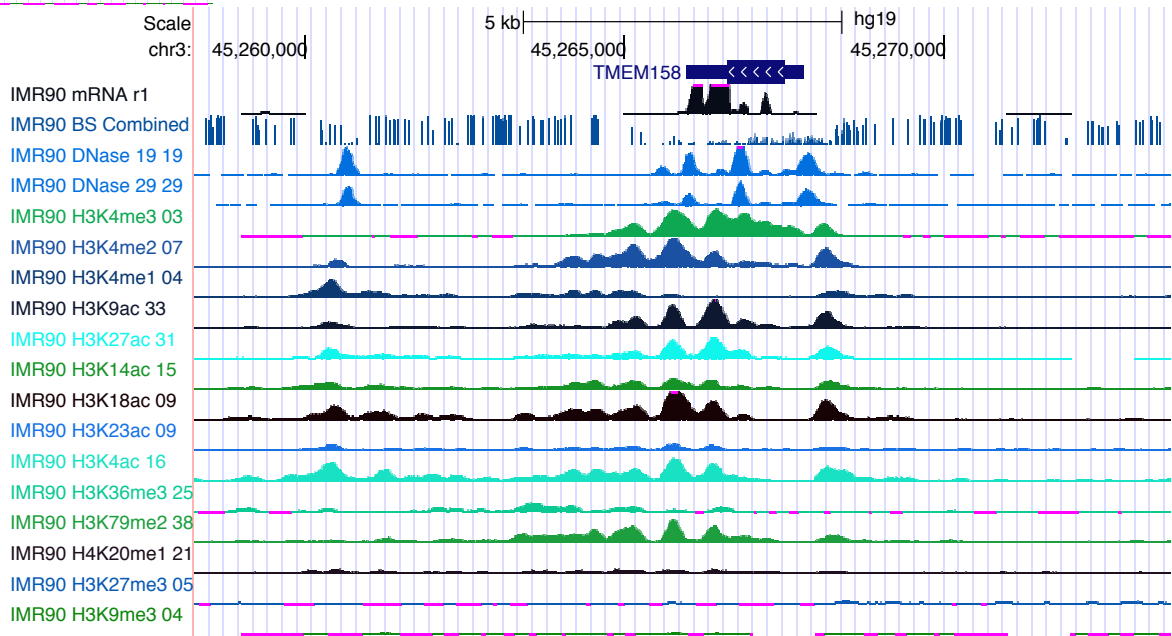
Wang et al., Combinatorial patterns of histone acetylations and methylations in the human genome, Nat Genet. 2008

Gene-Specific Associations among Histone Modifications



H3K79me3

H4K20me1



H3K79me3

H4K20me1

Talk Outline

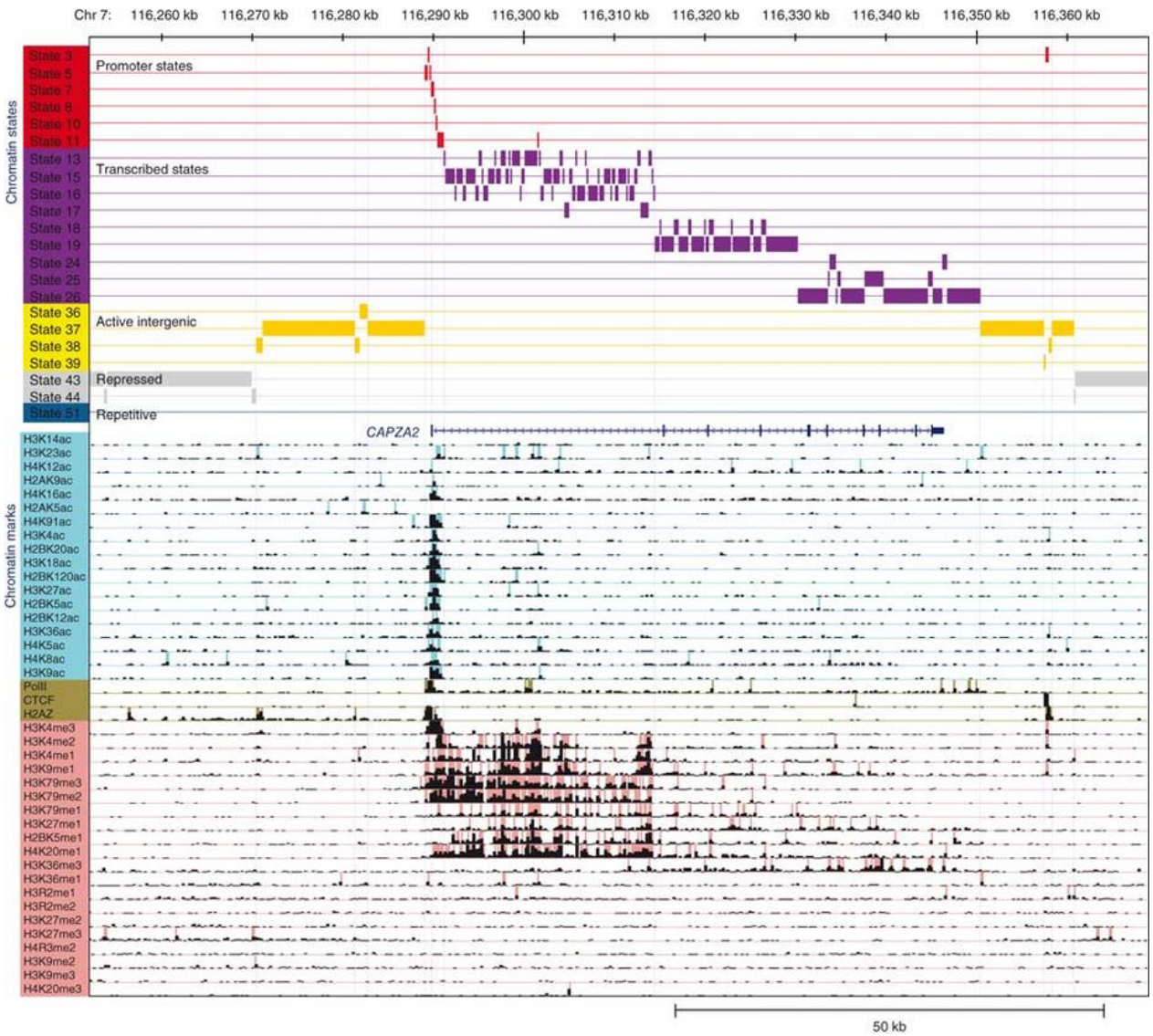
Understanding the Complexity of Histone Modifications

1. ChIP - Seq Analysis
2. Detecting Differential Modified Regions
3. Data Integration
4. **Identifying Chromatin States**

Future Directions

1. From Correlation to Causality
2. Single-Cell ChIP-Seq
3. Time Course Data

Ernst and Kellis, 2010



Talk Outline

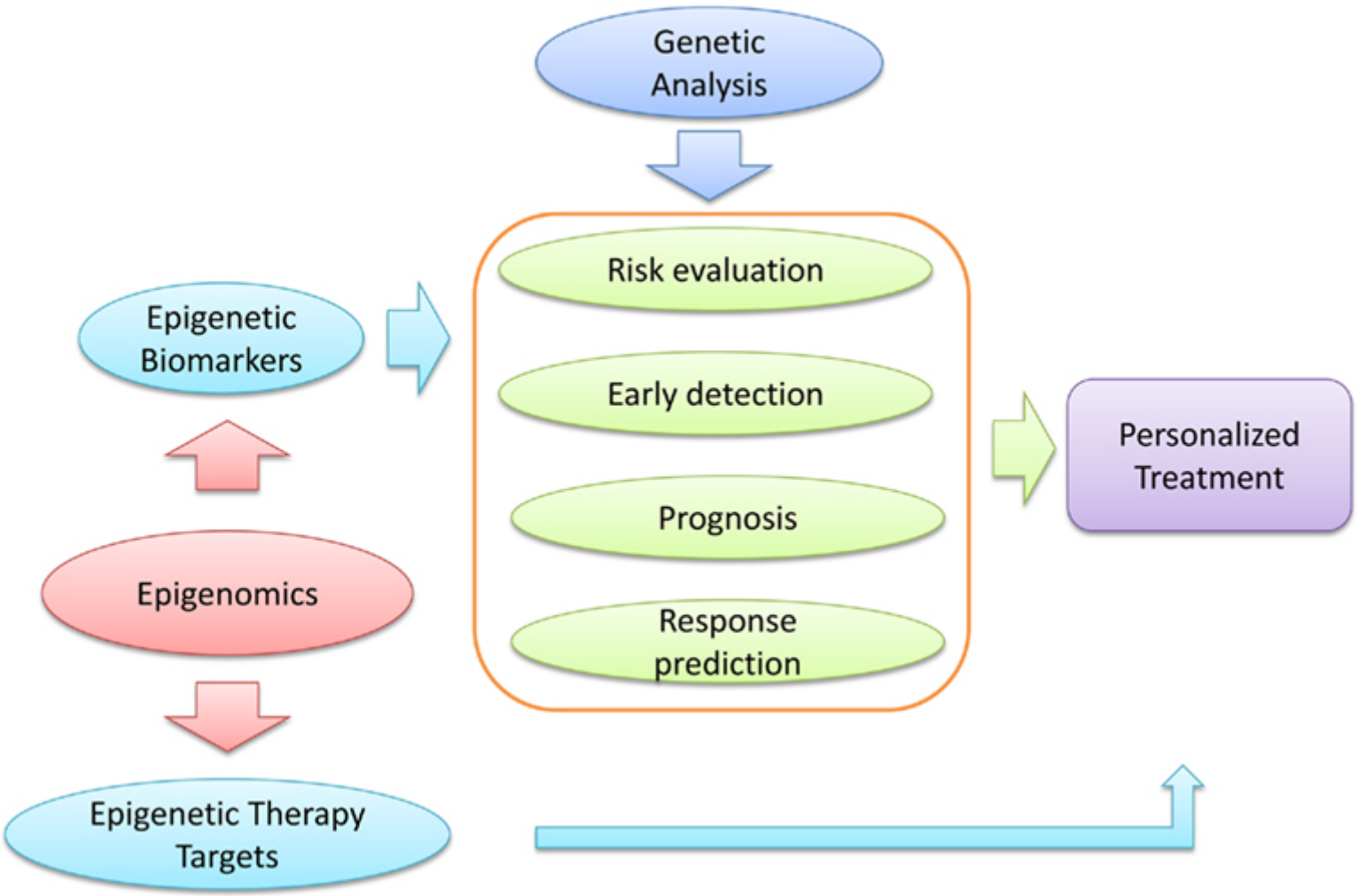
Understanding the Complexity of Histone Modifications

1. ChIP - Seq Analysis
2. Detecting Differential Modified Regions
3. Data Integration
4. Identifying Chromatin States

Future Directions

1. From Correlation to Causality
2. Single-Cell ChIP-Seq
3. Time Course Data

Epigenomics and Disease



ChIP-Seq Hands-On

Thanks

University of Edinburgh

- Prof. Guido Sanguinetti
- Prof. Adrian Bird

Medizinische Universitaet Wien

- Dr. Sabine Lagger
- Prof Christian Seiser

Imperial College London

- Saulius Lukauskas

Fondazione Bruno Kessler, Povo, Italy

- Roberto Visintainer

Funding: EU FP7 Marie Curie Actions / EMBO Long-Term Fellowship