Machine Learning in Bioinformatics

Introduction

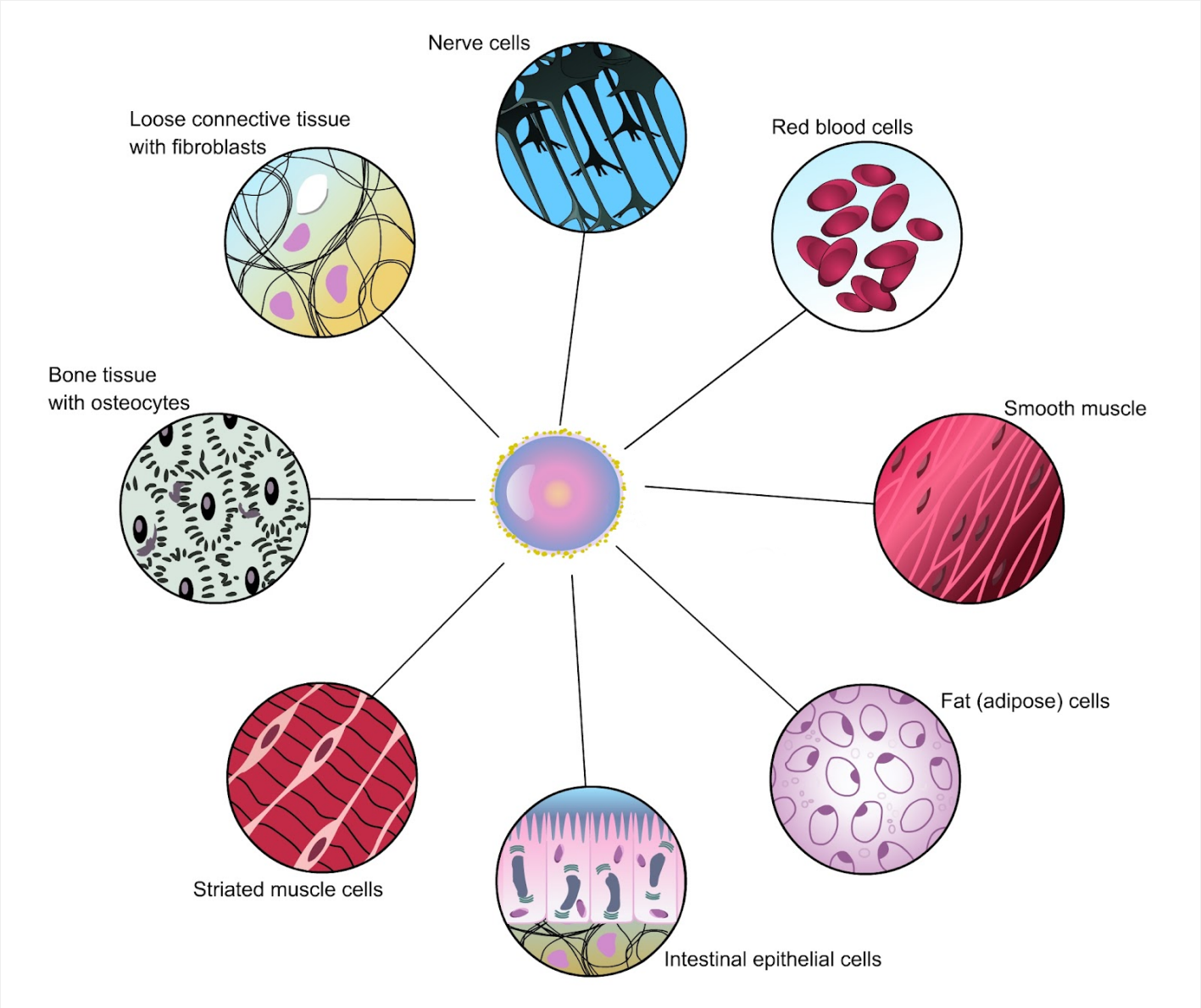
Philipp Benner
philipp.benner@bam.de

VP.1 - eScience
Federal Institute of Materials Research and Testing (BAM)

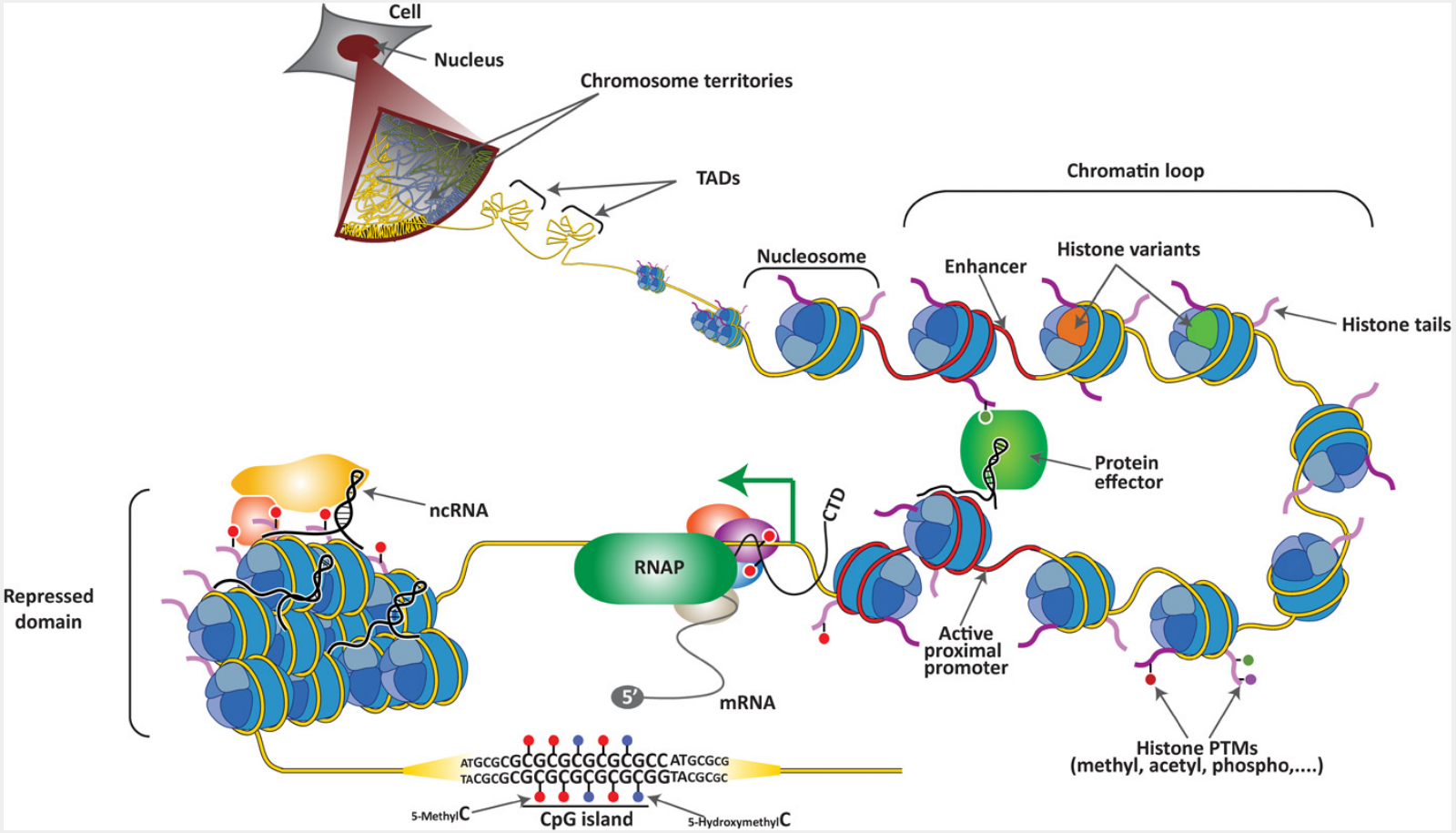
April 25, 2024

Biological Background

Biological background

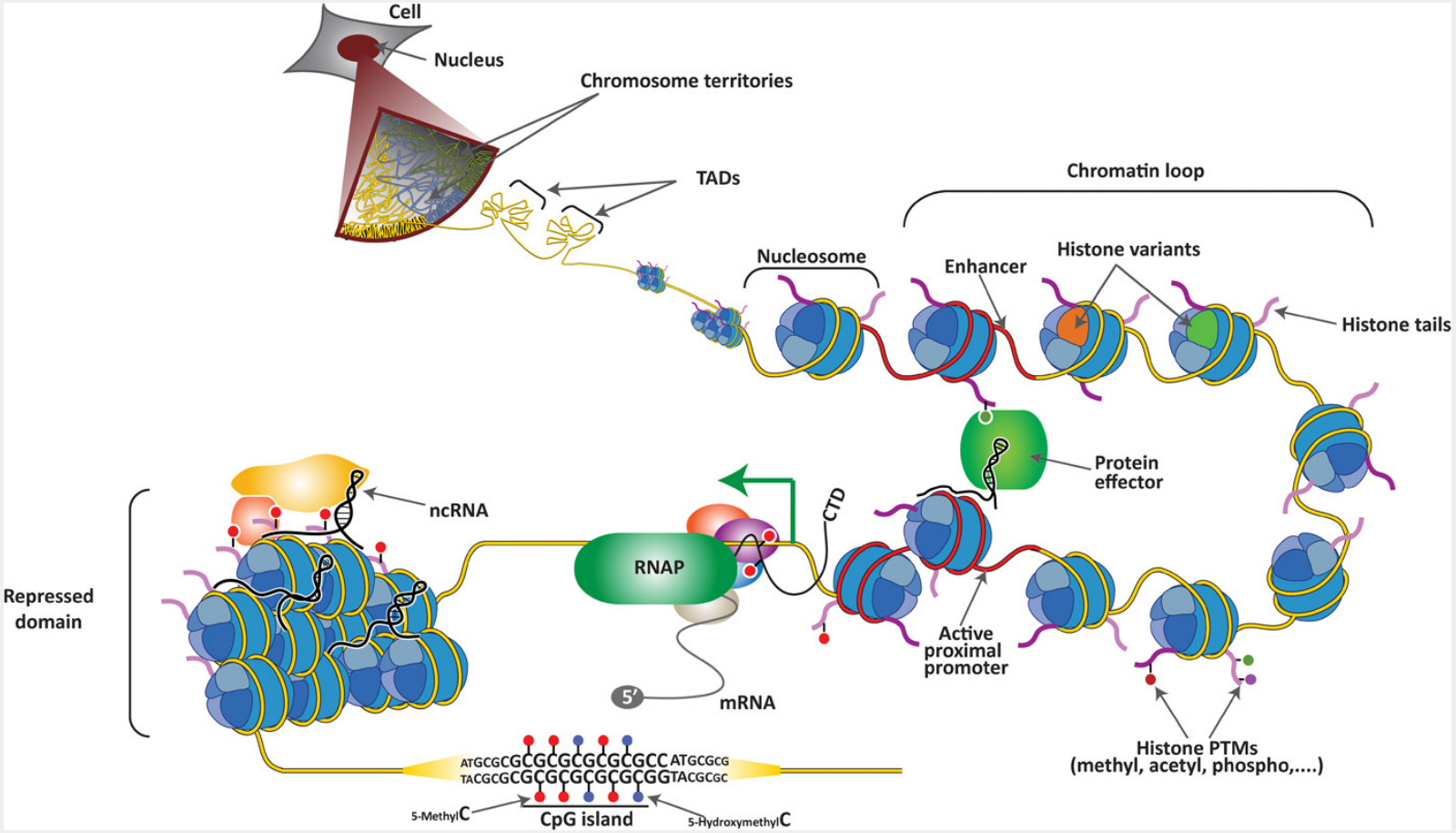


Biological background



[Aranda et al., 2015]

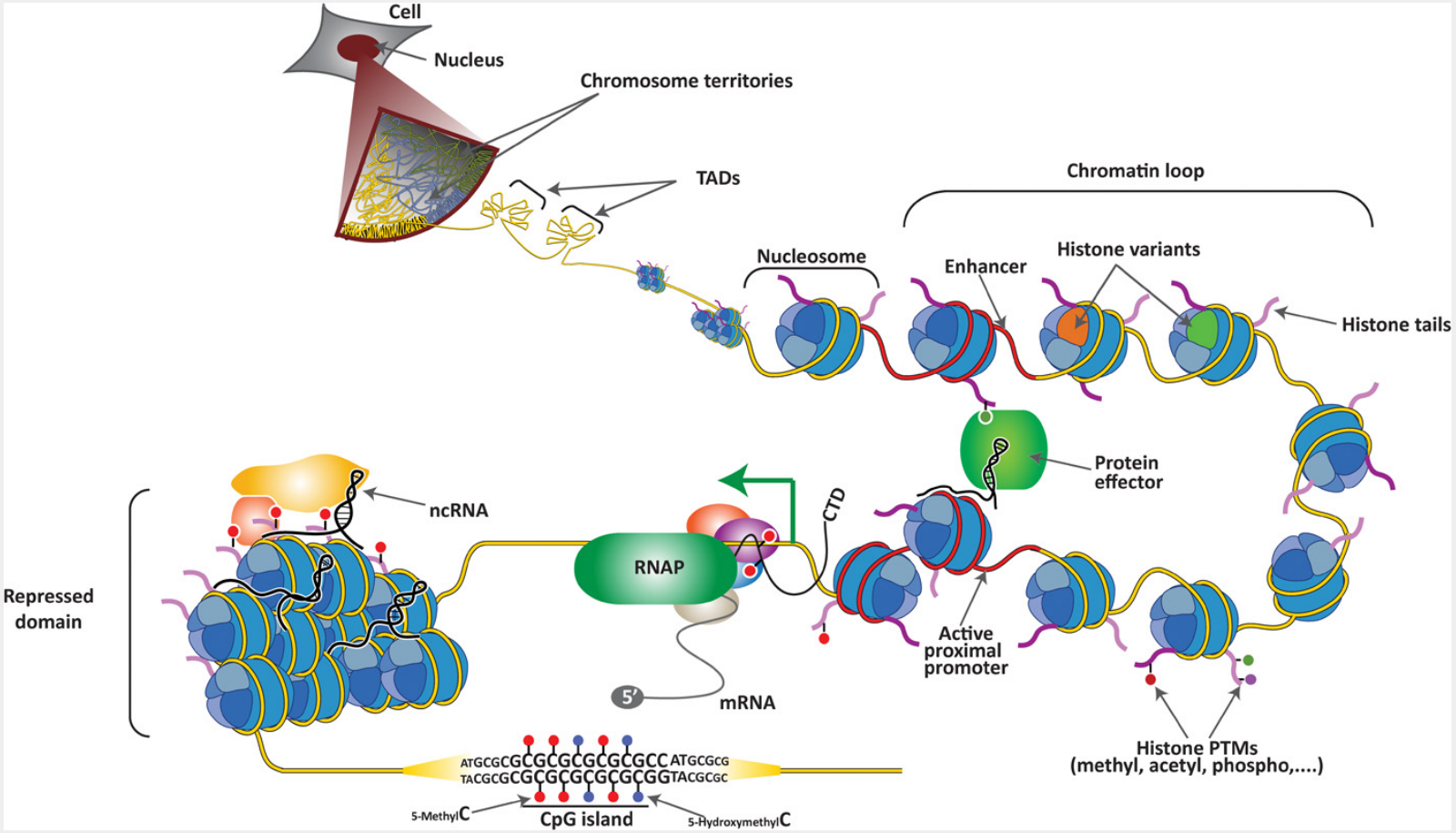
Biological background



[Aranda et al., 2015]

Khorana, Holley and Nirenberg (1953-1965): Discovery of the Genetic Code

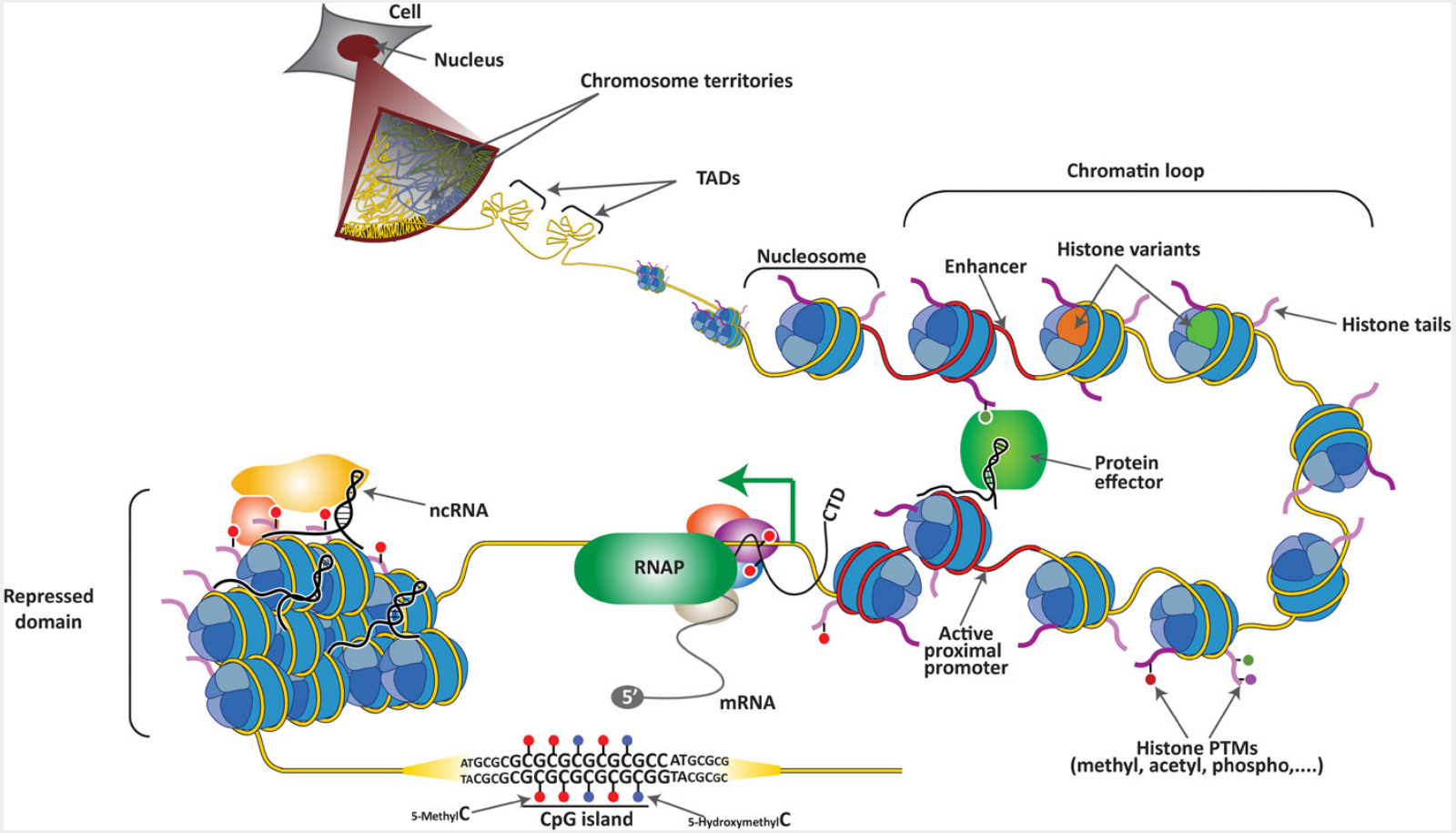
Biological background



[Aranda et al., 2015]

Human Genome Project (1990-2003): Identify DNA sequence (3 billion basepairs)

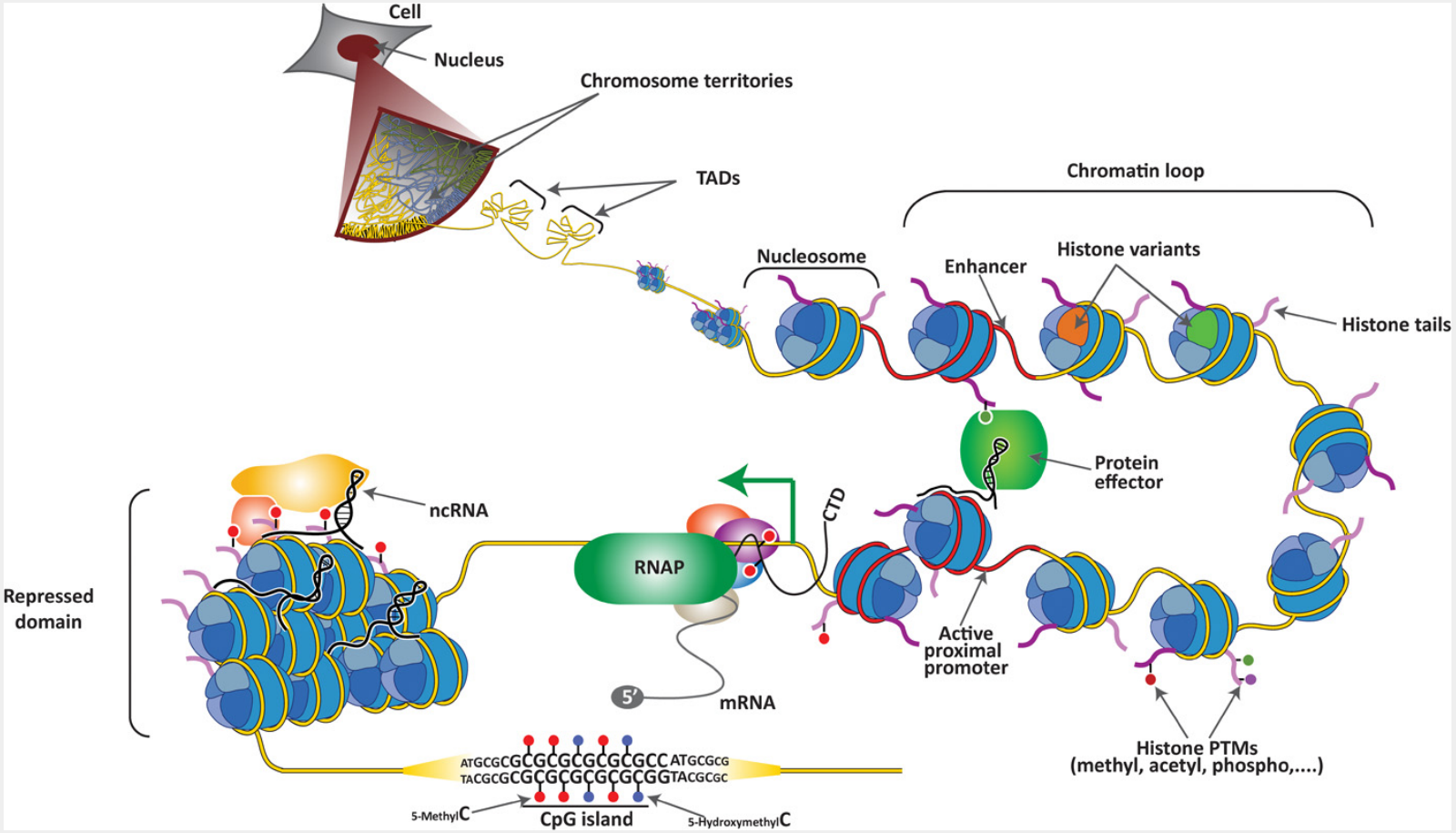
Biological background



[Aranda et al., 2015]

GENCODE Project (since 2003): Identify location of genes (20,000 protein coding genes)

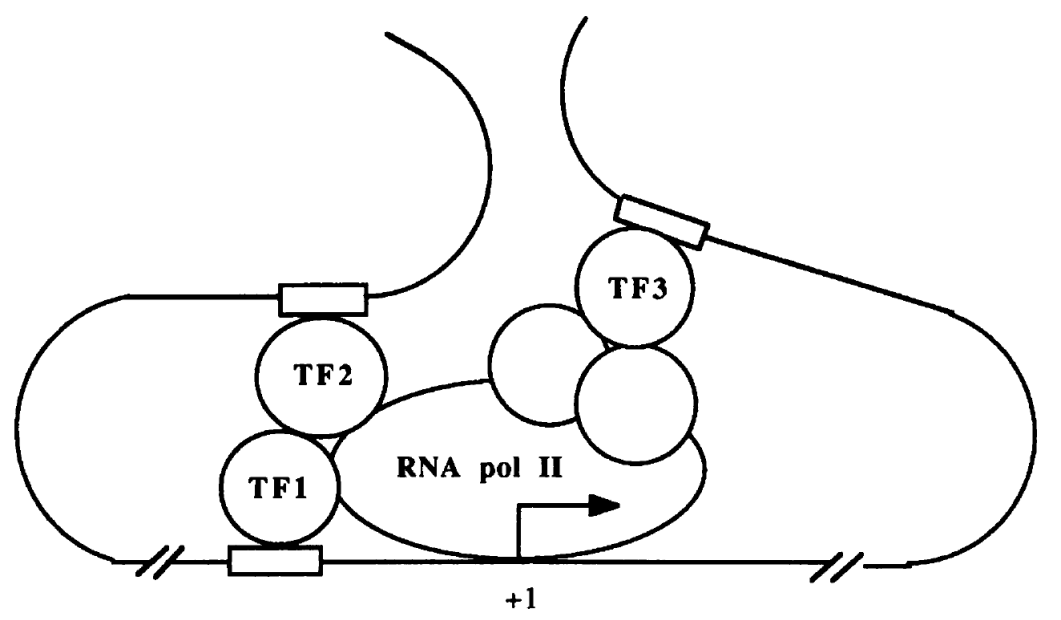
Biological background



[Aranda et al., 2015]

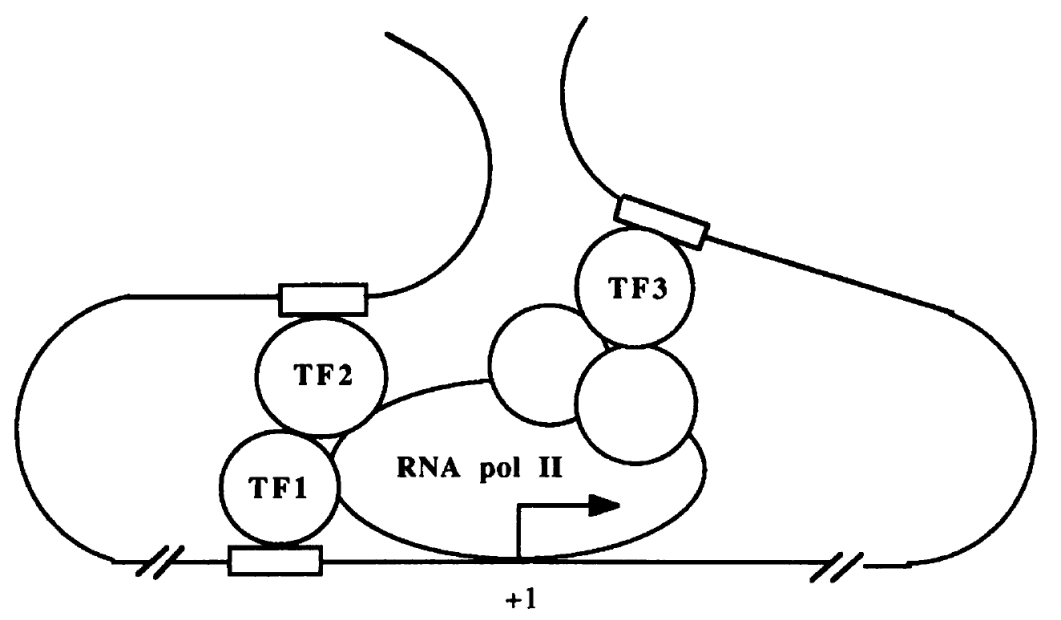
ENCODE Project (since 2003): Identify cell type-specific epigenetic marks

Gene Expression Regulation



[Mitchell and Tjian, 1989]

Gene Expression Regulation



[Mitchell and Tjian, 1989]

- Gene expression is regulated by promoters and enhancers
- Enhancer activity is highly cell type-specific
- Activation through transcription factors

Objectives

If we knew...

- all transcription factors
- their binding preferences
- and interactions
- all promoters
- all enhancers and their targets

we should be able to predict cell type-specific gene expression

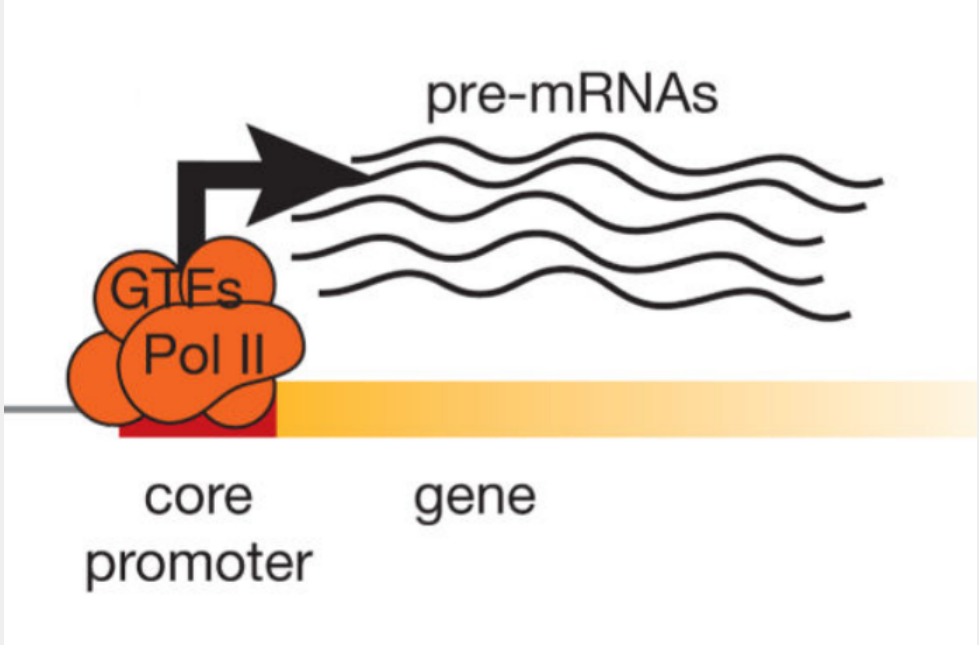
Application of machine learning in understanding gene regulation

- Study how the expression of genes is regulated
- How is formation about gene expression encoded in the DNA?
- How do promoters control gene expression?
- What is the role of enhancers?
- When do enhancers get active?
- How do enhancers link to promoters?

Applications of ML

Application 1: Promoter activity (regression)

How much control do promoters have over gene expression?



Approach: Develop machine learning method that predicts gene expression values from promoters

Application 1: Data

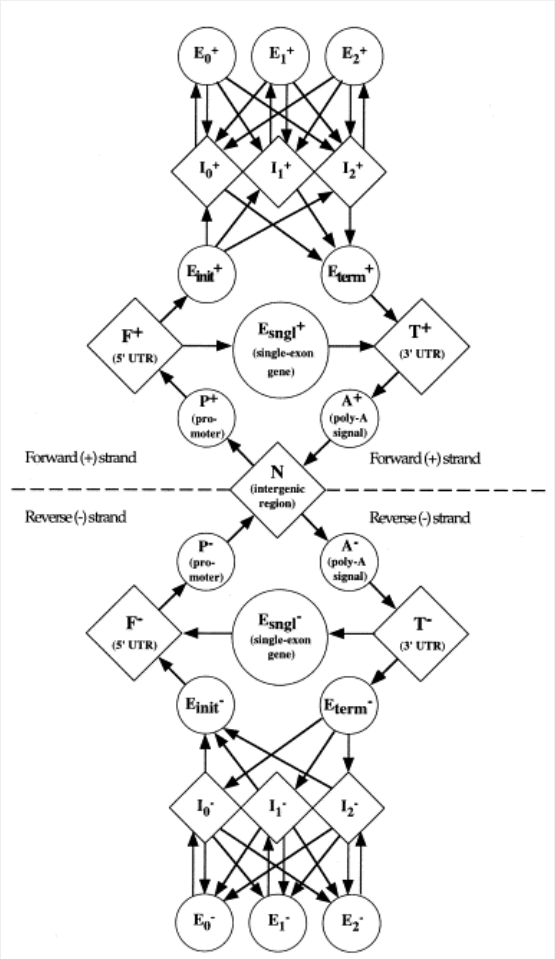
Expression values are cell type-specific. We only look at expression levels in liver:

Gene	Promoter sequence	Motif scores	Expression
1	CGTAGA...AGC	0.23,0.53,...,0.90	100
2	CTTGGA...CCC	0.03,0.87,...,0.93	328
...	...	...	...
N	GGACGA...AAT	0.69,0.21,...,0,43	0

Last column shows expression levels derived from total RNA-seq

Application 1: How do we know the position of genes?

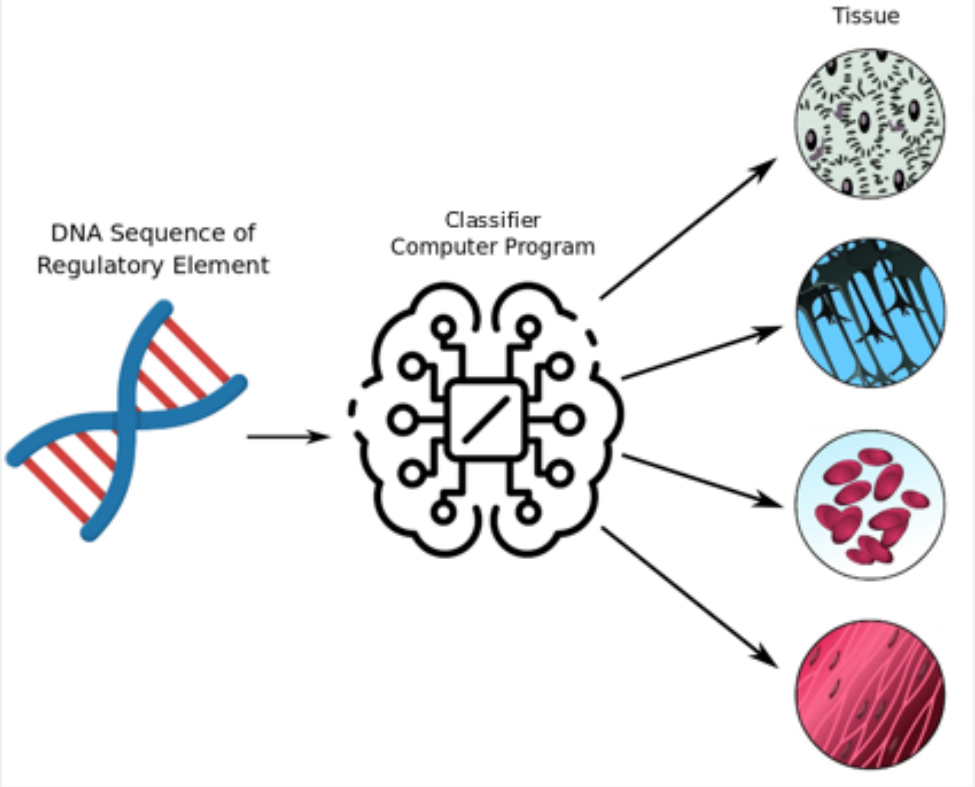
- Option 1: RNA-seq experiments of all tissues
- Option 2: Predictions from DNA-sequence
  - Hidden Markov model for gene structure prediction



[Burge and Karlin, 1997]

Application 2: Enhancer activity (classification)

Unterstand to what extent and how cell type-specific enhancer activity is encoded in the DNA sequence



Approach: Develop machine learning method that identifies relevant patterns

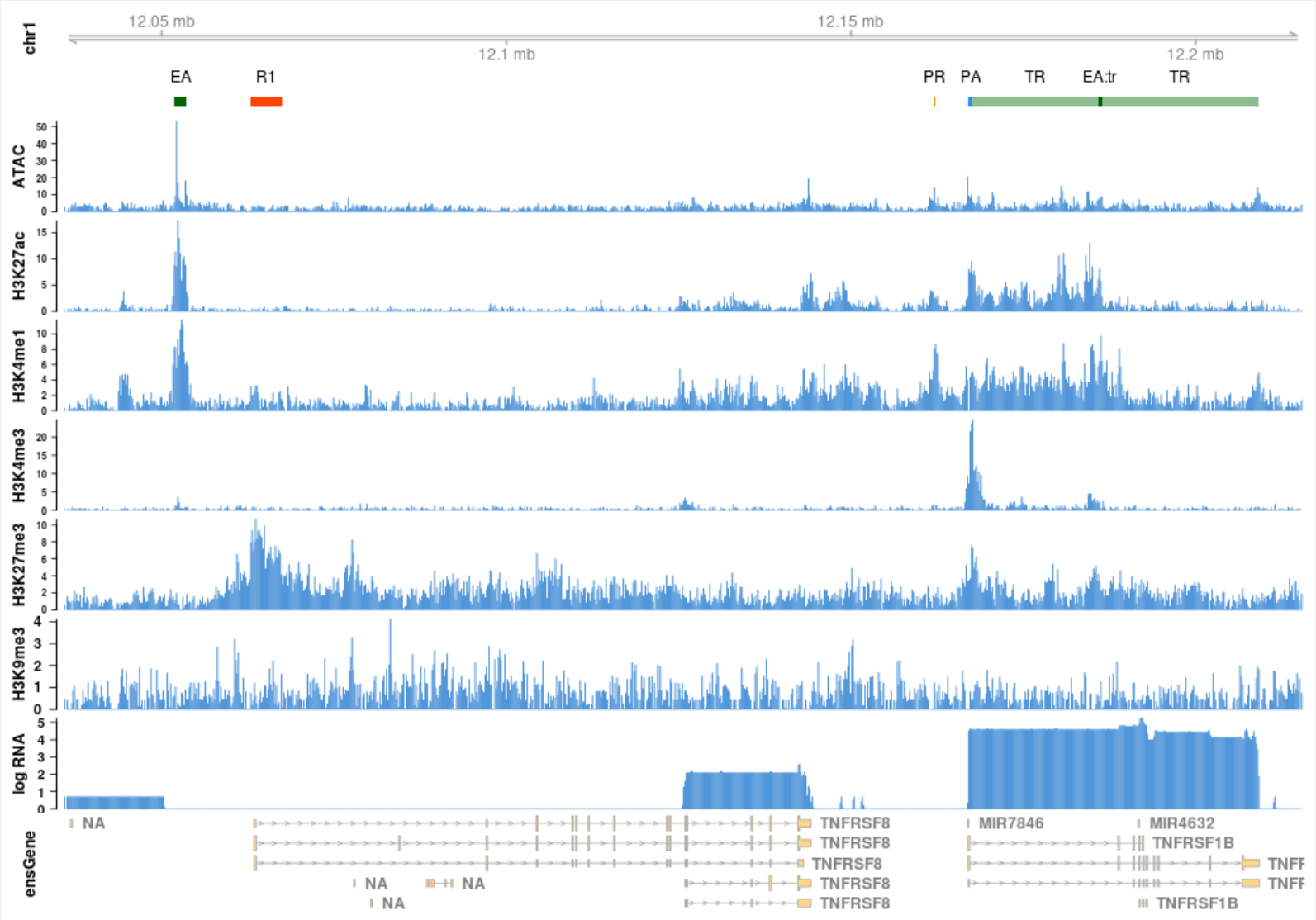
Application 2: Data

Enhancer activity is highly cell type-specific. We consider enhancers active in liver:

Enhancer	Sequence	Motif scores	Active
1	CGTAGA...AGC	0.23,0.53,...,0.90	1
2	CTTGGA...CCC	0.03,0.87,...,0.93	1
...	...	...	...
N	GGACGA...AAT	0.69,0.21,...,0,43	0

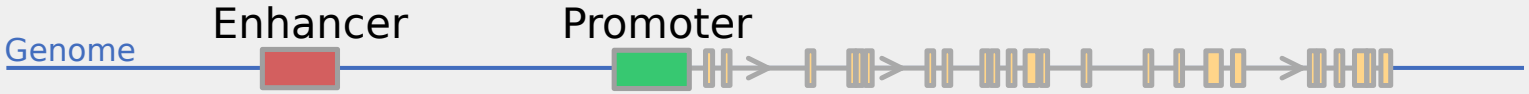
Last column encodes if an enhancer is active (1) or inactive (0) in liver

Identification of enhancers



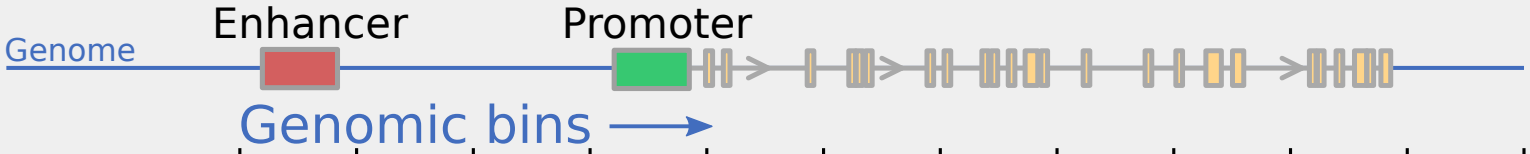
Identification of enhancers

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



Identification of enhancers

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



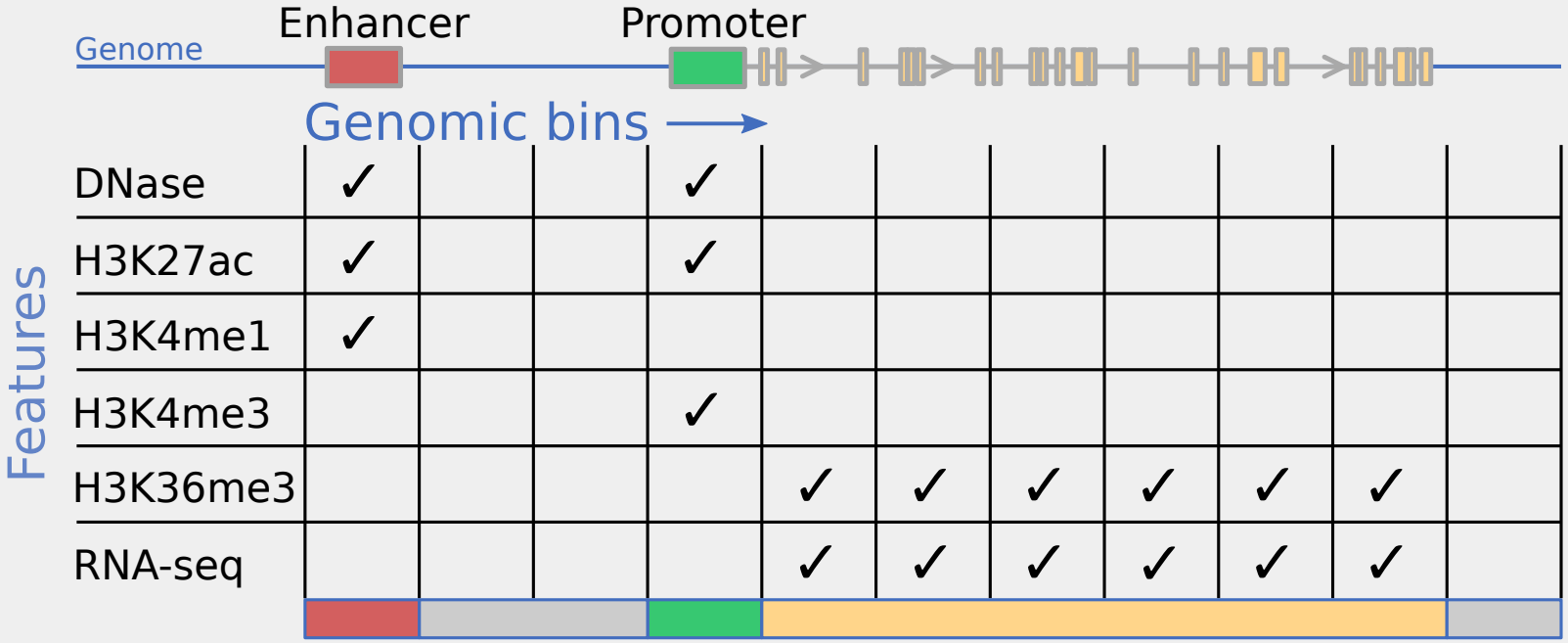
Genomic bins →

Features

DNase	✓			✓							
H3K27ac	✓			✓							
H3K4me1	✓										
H3K4me3				✓							
H3K36me3					✓	✓	✓	✓	✓	✓	
RNA-seq					✓	✓	✓	✓	✓	✓	

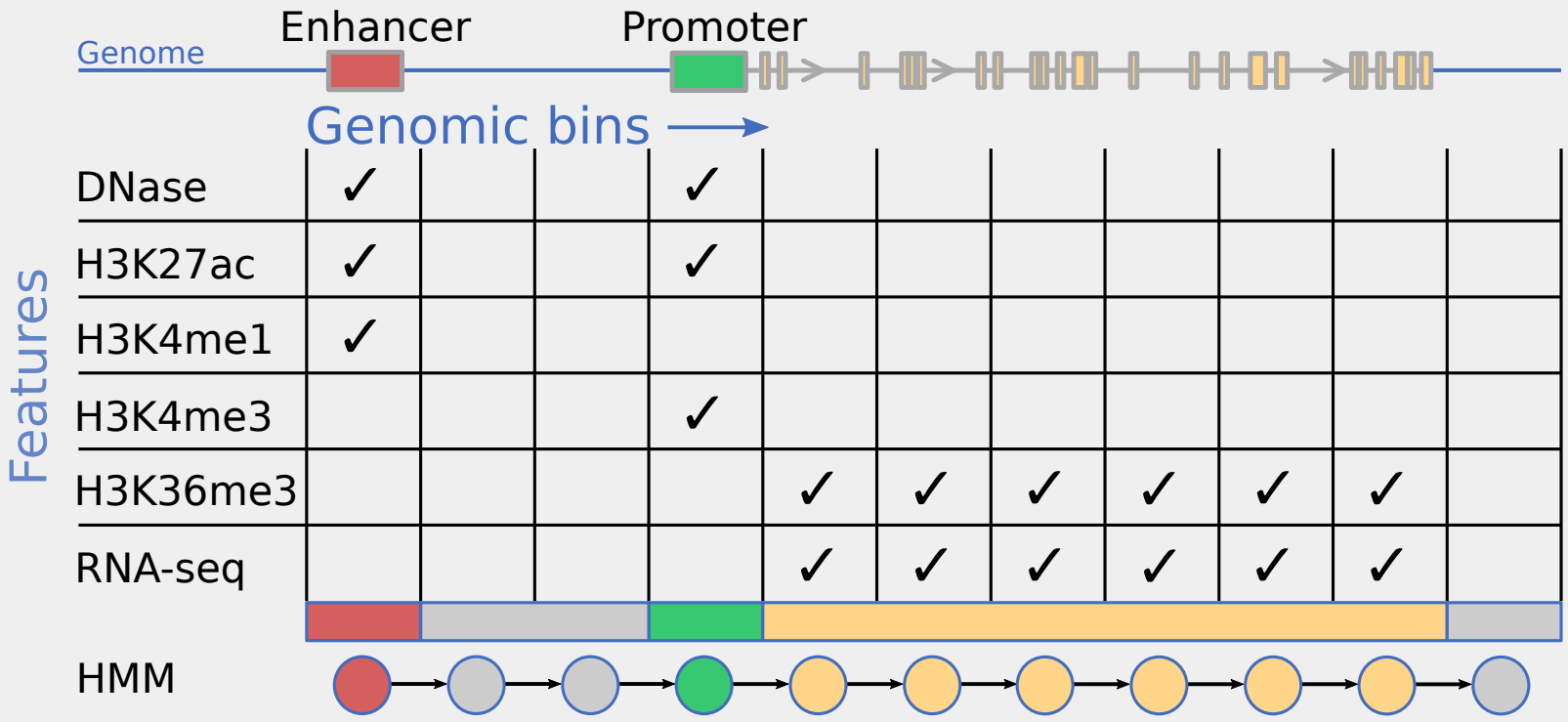
Identification of enhancers

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



Identification of enhancers

Enhancers are commonly identified from genome segmentations using hidden Markov models (HMMs):



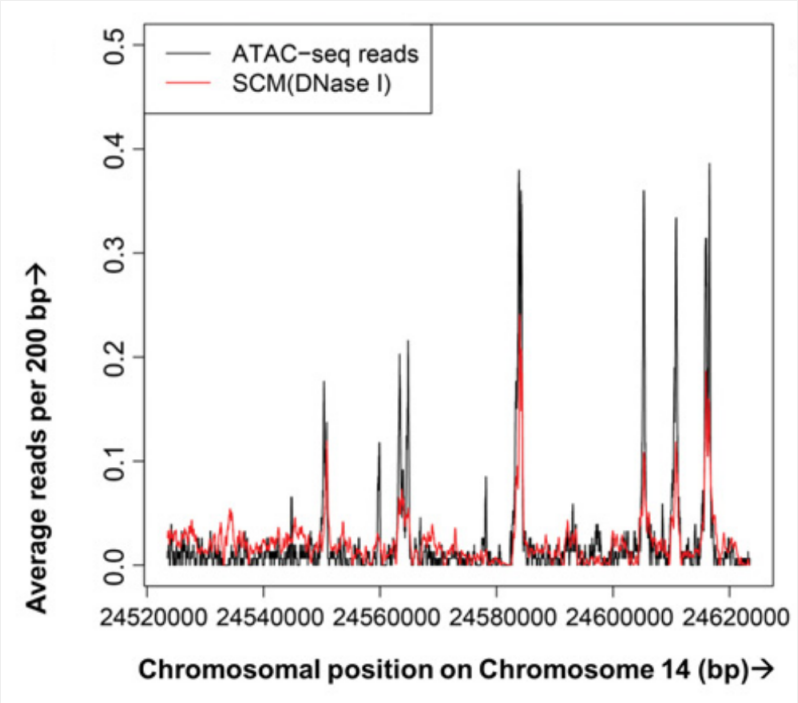
ENCODE Mouse Embryo Data

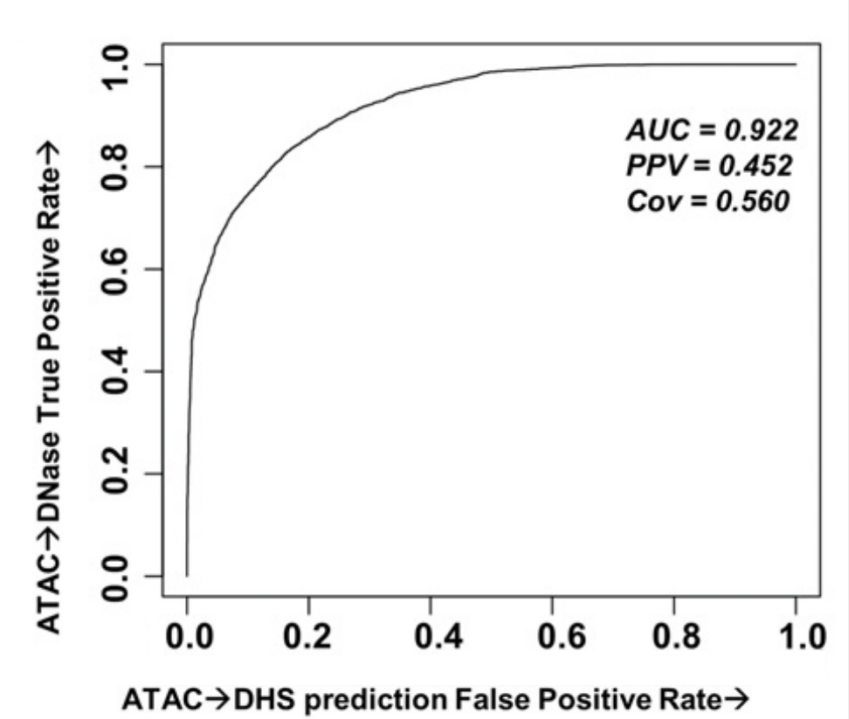
Enhancers are more difficult to identify. We need data from many different cell types:

	Forebrain	Midbrain	Hindbrain	Liver	Lung	Kidney	Heart	Limb
Day 11.5	✓	✓	✓	✓				
Day 12.5	✓	✓	✓	✓				
Day 13.5	✓	✓	✓	✓				
Day 14.5	✓	✓	✓	✓	✓	✓	✓	✓
Day 15.5	✓	✓	✓	✓	✓	✓	✓	✓
Day 16.5	✓	✓	✓	✓	✓	✓		

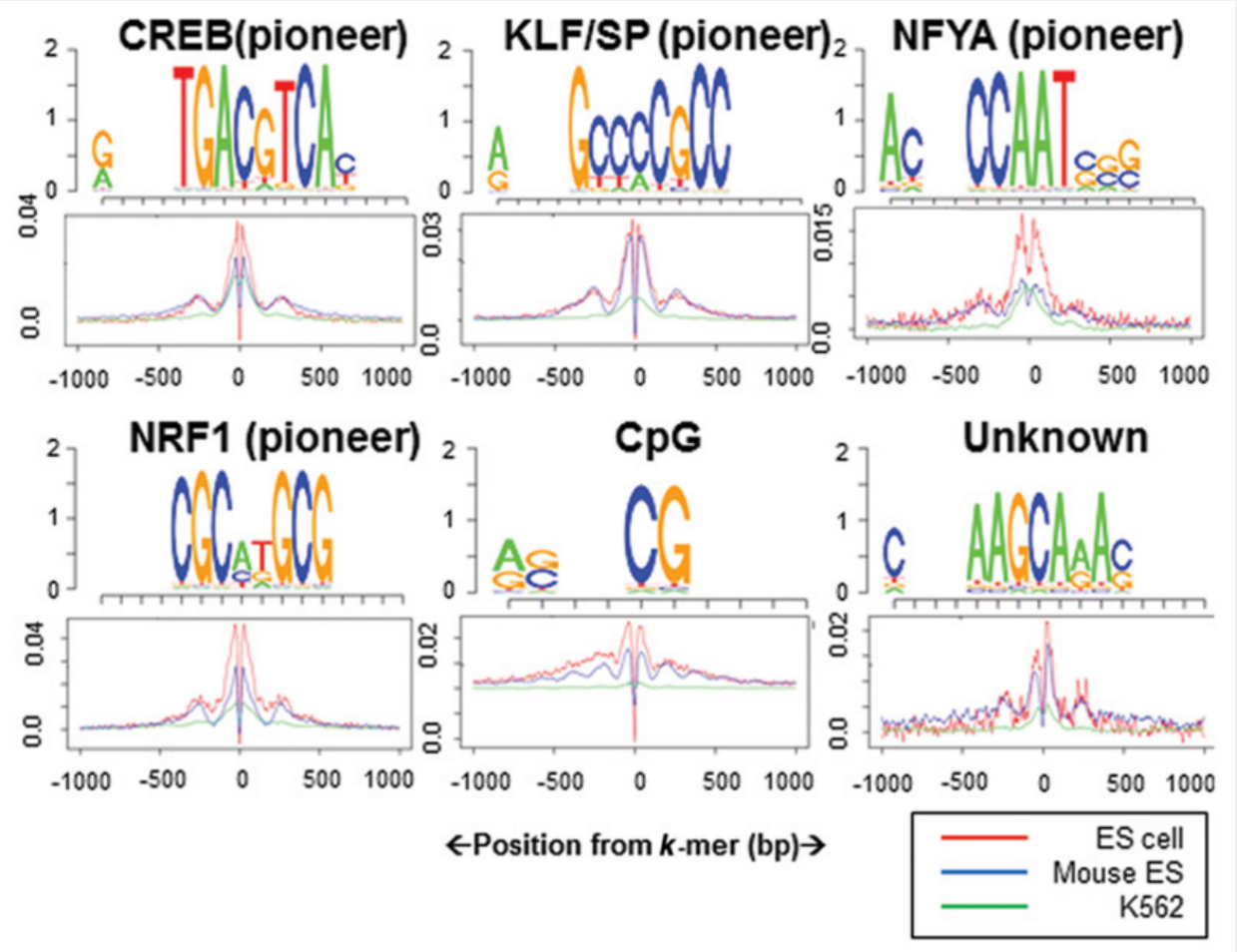
Related Studies

Prediction of Accessible Regions from DNA sequence [Hashimoto et al., 2016]

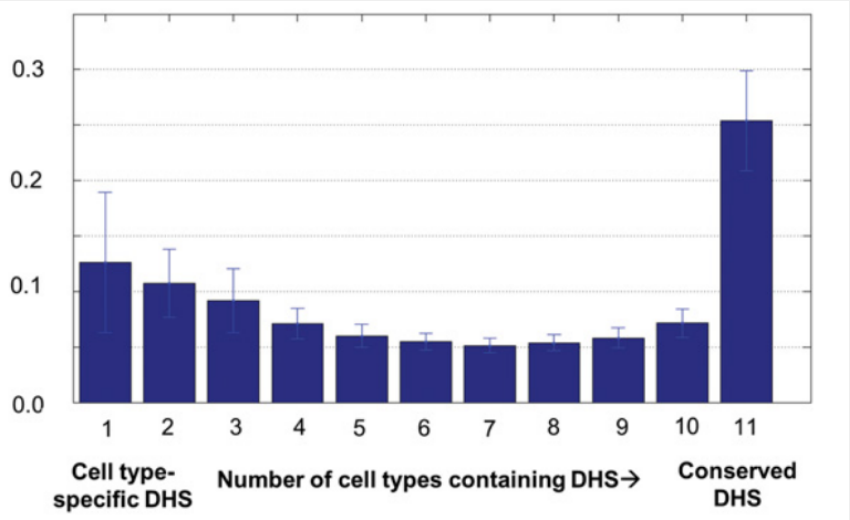


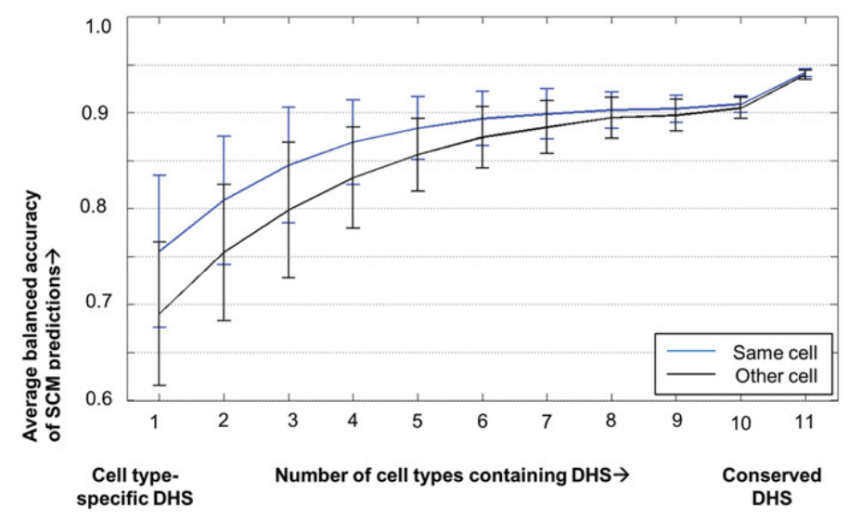


Prediction of Accessible Regions from DNA sequence [Hashimoto et al., 2016]



Prediction of Accessible Regions from DNA sequence [Hashimoto et al., 2016]





Software Requirements

Software requirements

- Python ($\geq$ 3.7) environment [Recommended: Anaconda]
- Python Packages:
  - Scikit-learn
  - Pandas
  - Numpy
  - PyTorch
- Editor for Jupyter Notebooks (e.g. VS Code)
- gkmSVM
  `https://www.beerlab.org/gkmsvm/`

References I

Aranda, S., Mas, G., and Di Croce, L. (2015).
Regulation of gene transcription by polycomb proteins.
Science advances, 1(11):e1500737.

Burge, C. and Karlin, S. (1997).
Prediction of complete gene structures in human genomic dna.
Journal of molecular biology, 268(1):78–94.

Hashimoto, T., Sherwood, R. I., Kang, D. D., Rajagopal, N., Barkal, A. A., Zeng, H., Emons, B. J., Srinivasan, S., Jaakkola, T., and Gifford, D. K. (2016).
A synergistic dna logic predicts genome-wide chromatin accessibility.
Genome research, 26(10):1430–1440.

Mitchell, P. J. and Tjian, R. (1989).
Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins.
Science, 245(4916):371–378.