University of Trieste
Master in Data Science and Scientific Computing
Master Thesis

An information-theoretic investigation into epigenetic regulation of gene expression

DATA SCIENCE &
SCIENTIFIC COMPUTING

Candidate:
Davide Scassola

Supervisors:
Guido Sanguinetti,
Matteo Marsili

Academic year 2019/2020

Outline

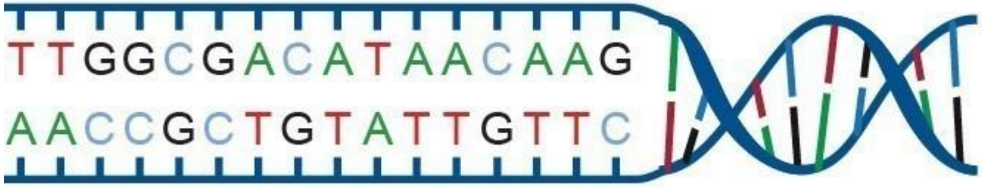
1. Epigenetics and methylation

2. Application of *Multi-Scale Relevance* to Methylation Data

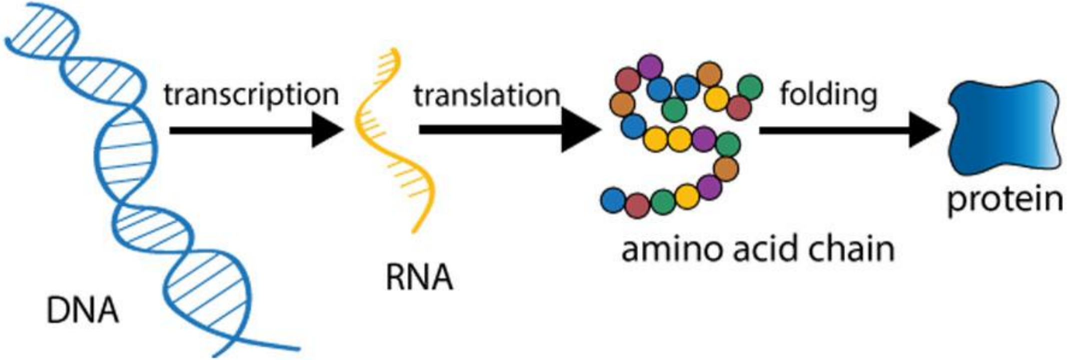
3. Relationship with Gene Expression

Genetics Recap

In each cell of an individual the same copy of the genetic information is stored in DNA

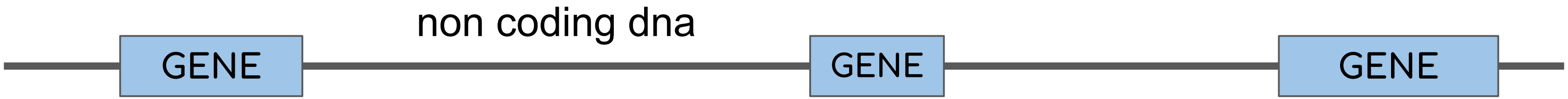


DNA encodes information for the synthesis of useful molecules: RNAs and proteins



Genetics Recap

- About 98,5% of the genome does not encode proteins
- The remaining regions are the genes



non coding dna

GENE

GENE

GENE

- non coding dna can have a function

Epigenetics

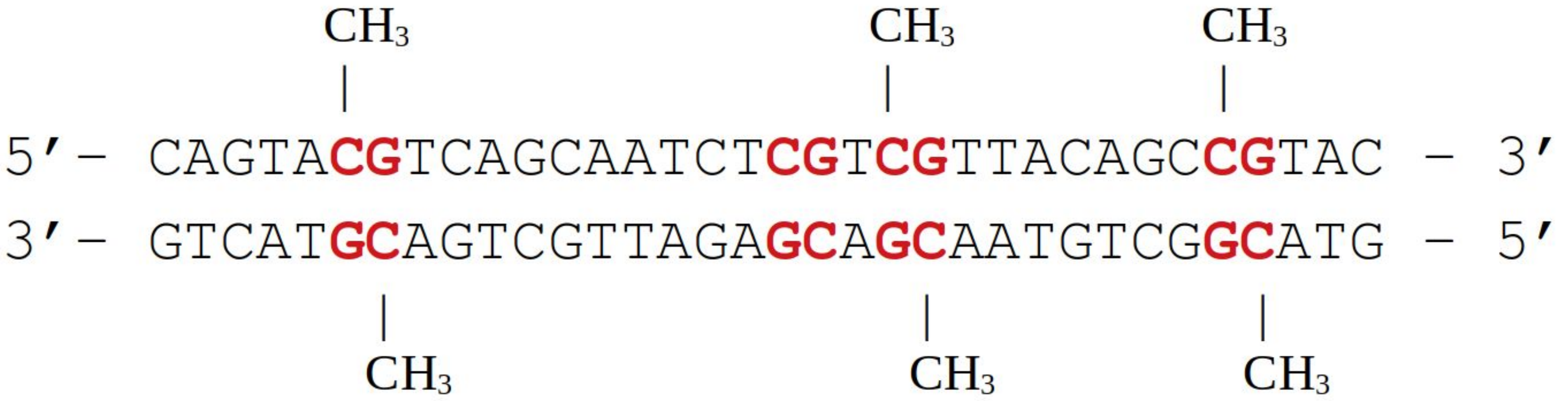
- Cells have the same DNA but express genes differently

muscle cells
liver cells
nerve cells

- **Epigenetics**: heritable molecular changes that do not involve DNA base sequence

DNA Methylation

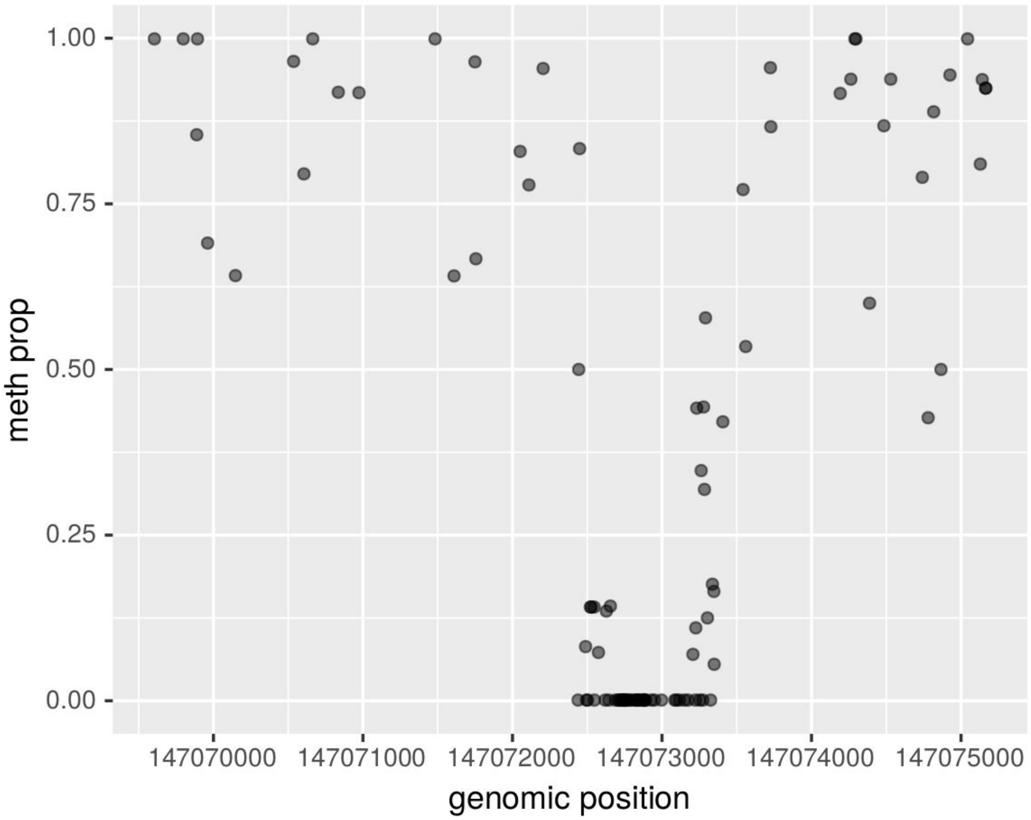
- It's the addition of a methyl group to a base
- In humans it mainly involves cytosines of CpG dinucleotides



5′ – CAGTA**CG**TCAGCAATCT**CG**T**CG**TTACAGC**CG**TAC – 3′

3′ – GTCAT**GC**AGTCGTTAGA**GC**A**GC**AATGTCG**GC**ATG – 5′

Methylation Data



- “Averaged” data from a sample of cells

DNA Methylation regulatory role

- Poor understanding of its influence on gene expression
- Common practice is to analyze mean methylation level for a region
- Just a slight negative correlation with gene-expression "genome-wide"
- Recent studies are focusing on more complex features (Kapourani and Sanguinetti 2016)

Problem statement

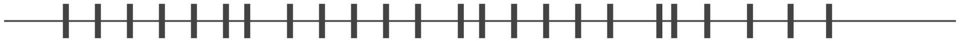
How much does methylation influences gene expression?

Does methylation patterns encode useful information?

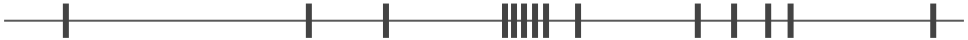
We explored the application of MSR to dna methylation data

Multi Scale Relevance

- MSR is a recently developed statistic: $\mathbb{R}^{\mathrm{n}} \rightarrow \mathbb{R}$
- Motivations rooted in Information Theory
- measure of “information” based on richness of density of states at different scales

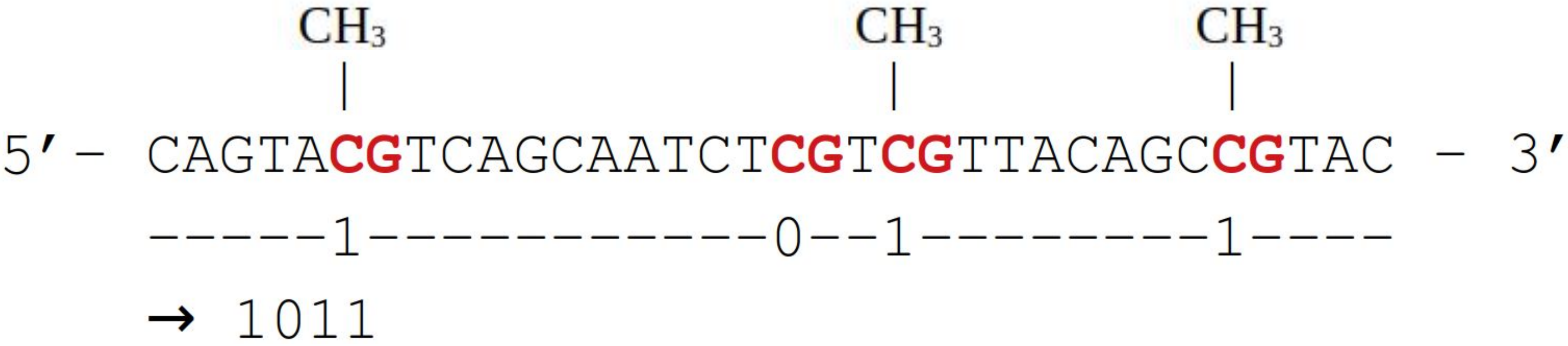


Low MSR
(“low information content”)



Higher MSR
(“higher information content”)

MSR on Methylation Data



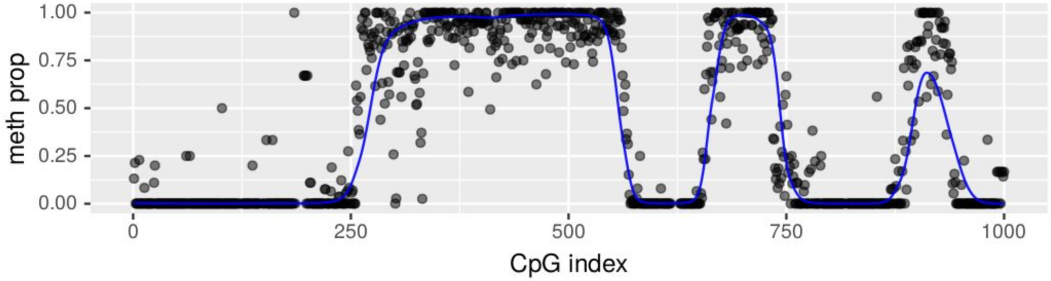
5′ – CAGTA**CG**TCAGCAATCT**CG**T**CG**TTACAGC**CG**TAC – 3′

-----1-----------0--1---------1----

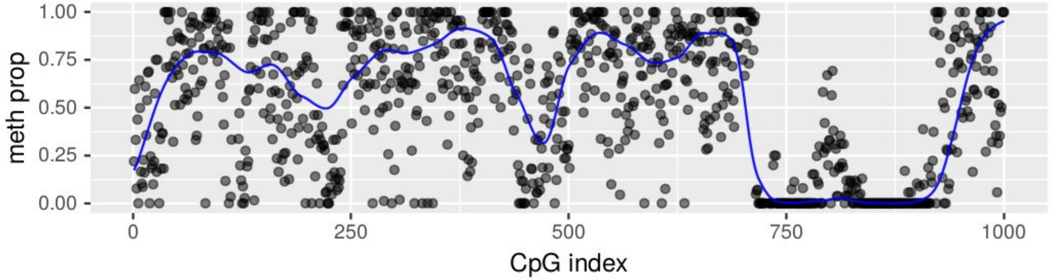
→ 1011

→ MSR on indexes of methylated (or unmethylated) CpGs

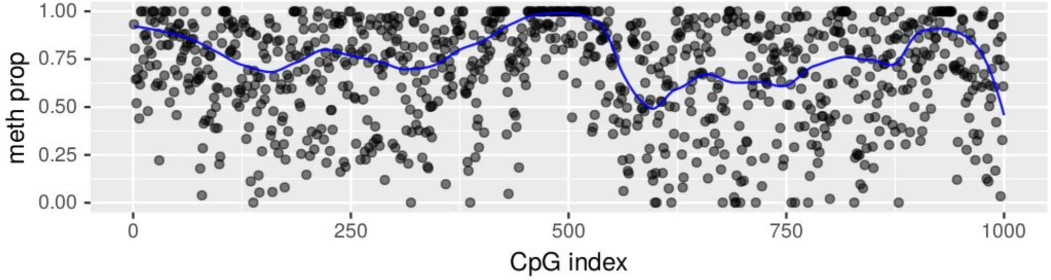
Genome-wide application



Low MSR



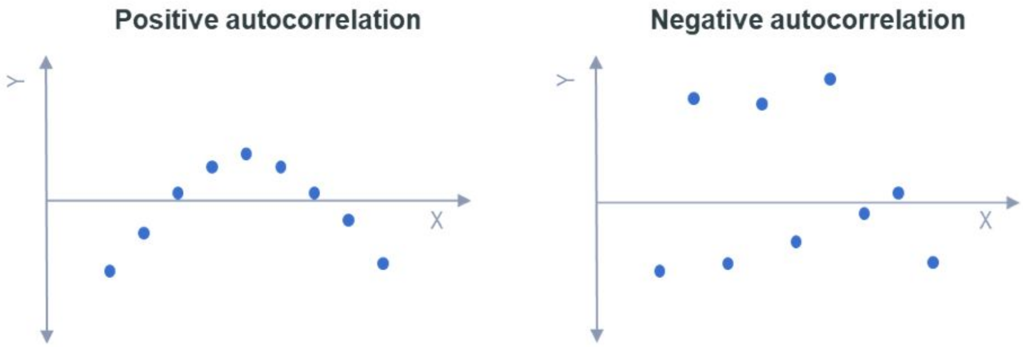
Medium MSR



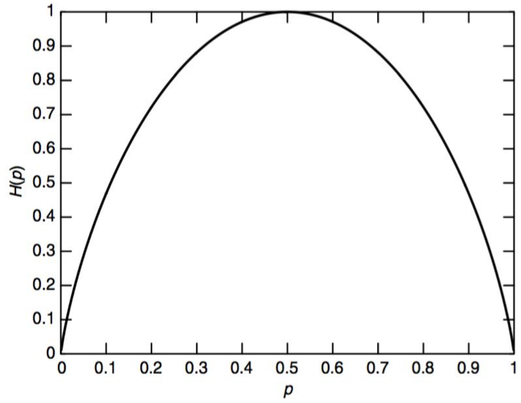
High MSR

↓MSR

↑autocorrelation

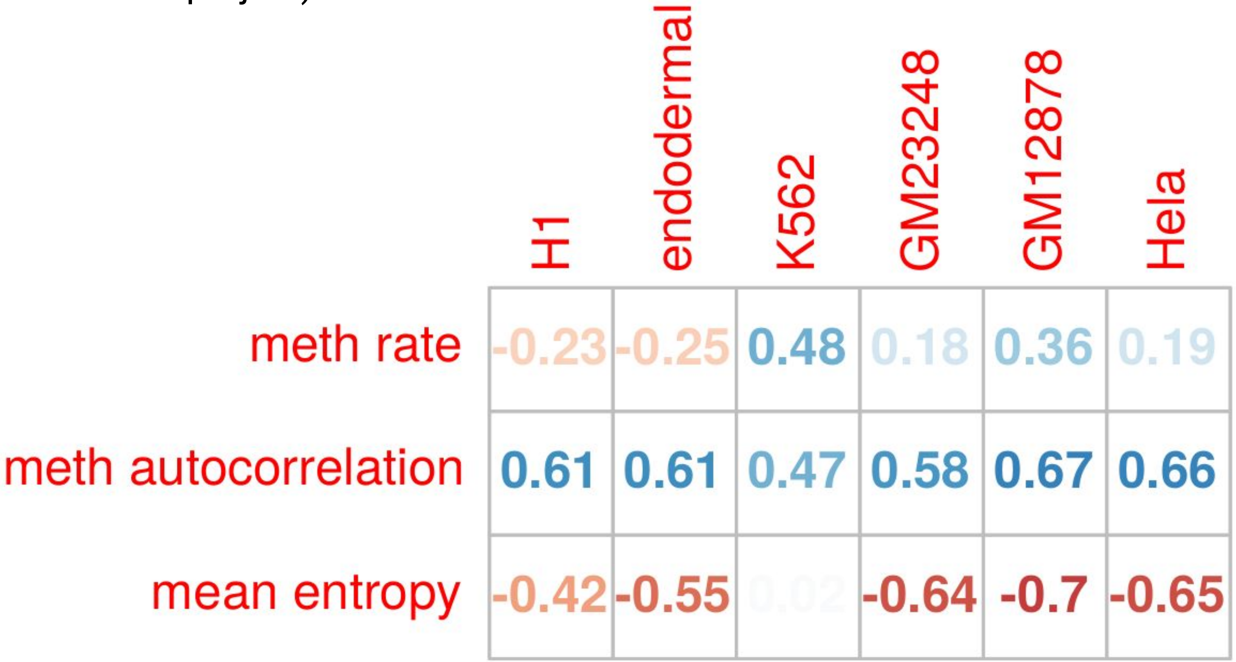


↓entropy



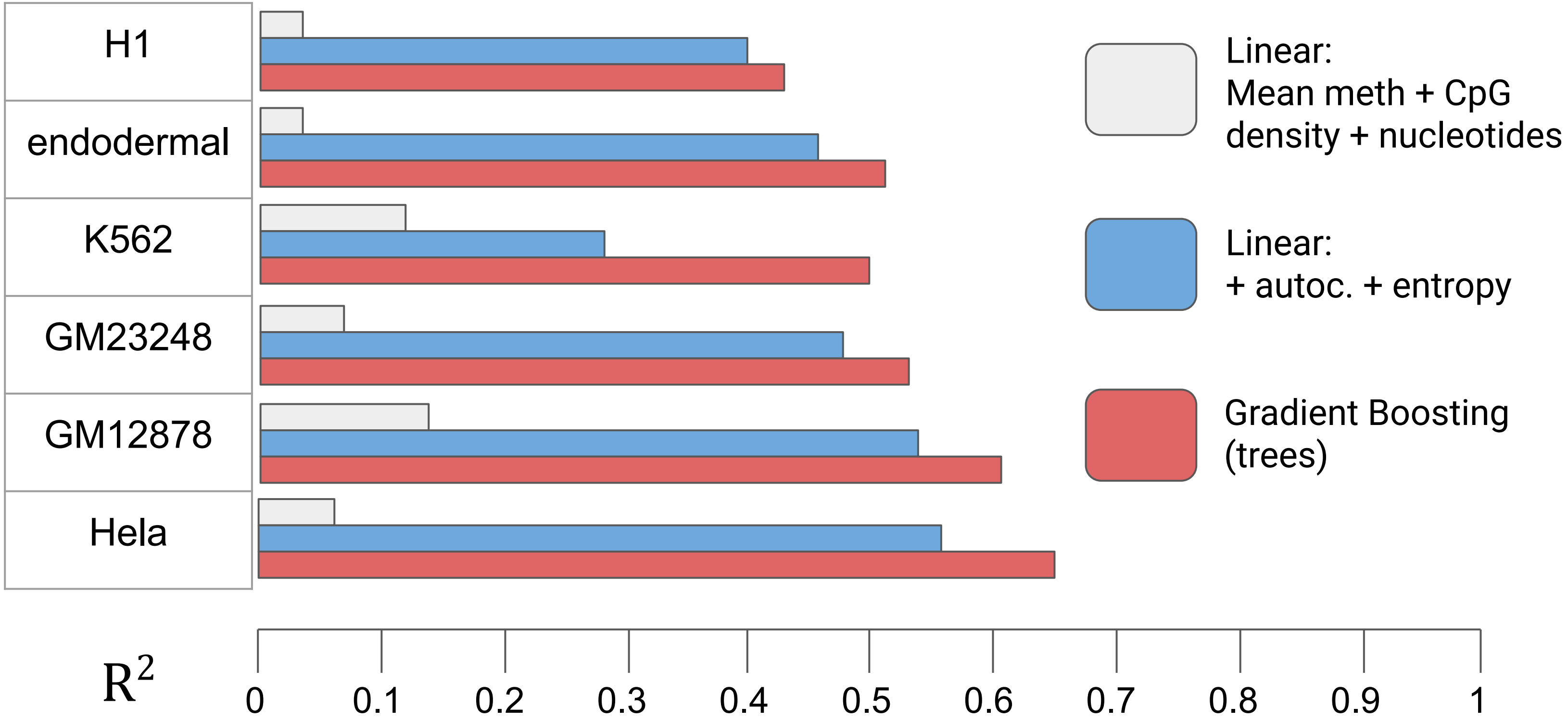
Correlation with expression (at gene bodies)

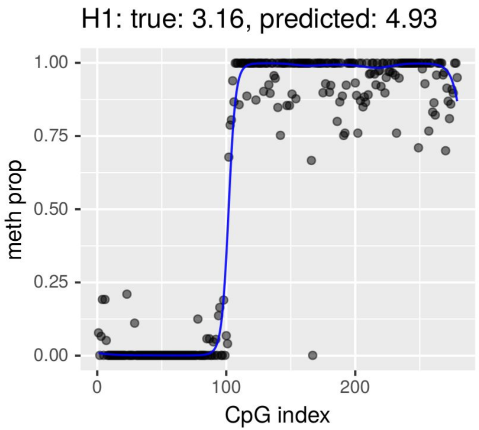
(data from ENCODE project)

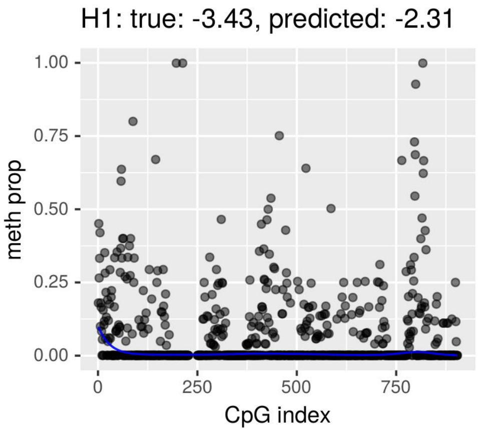


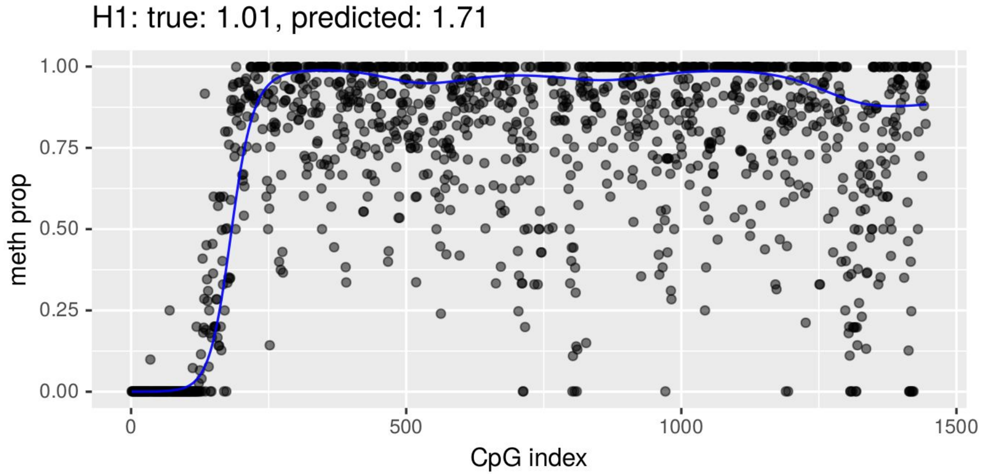
	H1	endodermal	K562	GM23248	GM12878	Hela
meth rate	-0.23	-0.25	0.48	0.18	0.36	0.19
meth autocorrelation	0.61	0.61	0.47	0.58	0.67	0.66
mean entropy	-0.42	-0.55	0.02	-0.64	-0.7	-0.65

Models (for gene bodies)









Expression is generally higher where:

- Neat separation between methylated and unmethylated regions
- Homogeneity between cells
- Coexistence of methylated and unmethylated regions

↓

Coherent with recent research

Results

- We found methylation characteristics that seem determinant in gene expression

- Considerable improvement with respect to the model that only consider mean methylation level and CpG density

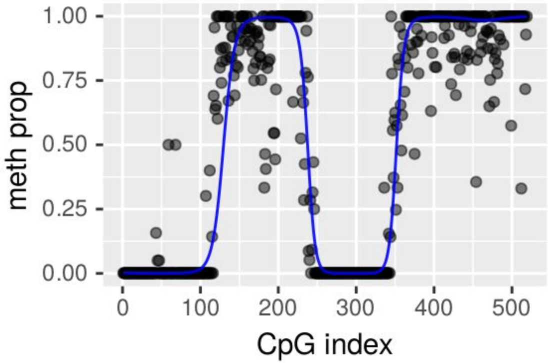
- Models hold for arbitrary regions

Limits

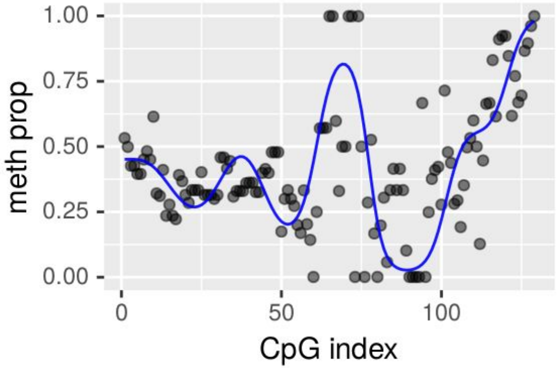
- Focus on correct functional region could be needed

- Not purely epigenetic study

- This application of MSR was not more useful in prediction than some a posteriori extracted features



H1
true: -3.43, predicted: 4.06



H1
true: 6.74, predicted: -3.73

Possible improvements

- Consider larger areas around genes
- More detailed description
- Investigate the role of these features in tissue-specific genes
- Apply MSR in a different way (genomic positions)

Thank you!

[1] Moore, L. D., Le, T., and Fan, G. (2013). Dna methylation and its basic function.

[2] Cubero, R. J., Marsili, M., and Roudi, Y. (2020). Multiscale relevance and informative encoding in neuronal spike trains.

[3] Cubero, R. J., Jo, J., Marsili, M., Roudi, Y., and Song, J. (2019). Statistical criticality arises in most informative representations.

[4] Marsili, M., Mastromatteo, I., and Roudi, Y. (2013). On sampling and modeling complex systems

[5] Kapourani, C.-A. and Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies.

MIROSOME, Luisa Lente