An Information-Theoretic Investigation Into Epigenetic Regulation of Gene Expression

Davide Scassola

Supervisors: Guido Sanguinetti, Matteo Marsili

University of Trieste
Department of Mathematics and Geosciences
Master in Data Science and Scientific Computing

December 9, 2020

Table of Contents

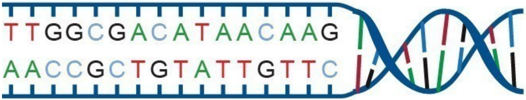
1. What is epigenetics and methylation
2. Application of MSR to Methylation Data
3. Relationship with Gene Expression

Table of Contents

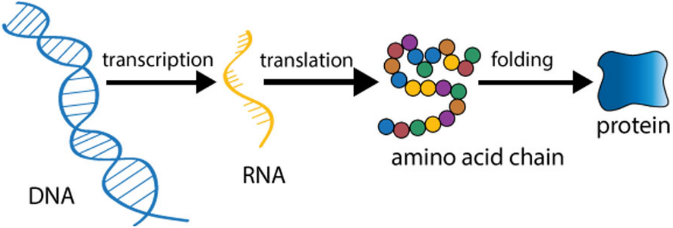
1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

Genetics

- In an individual each cell contains the same copy of the genetic information in DNA.

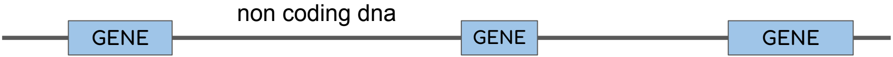


- DNA encodes information for the synthesis of useful molecules: RNAs and proteins.



Genetics

- About 98.5% of the genome does not encode proteins (non-coding DNA). The remaining regions are the genes.



- Non coding DNA regions can have a function.

Table of Contents

1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

Epigenetics

- Genes are not expressed in the same way in all cells, despite cells have (almost) the same genome.

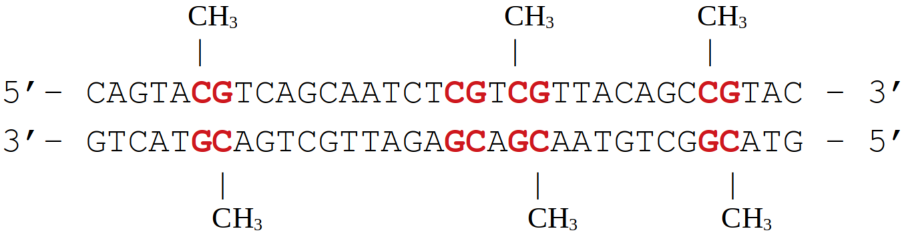
Epigenetics

- Genes are not expressed in the same way in all cells, despite cells have (almost) the same genome.
- Epigenetics is the study of the molecular mechanisms that involve DNA without altering its base sequence, that influence the **phenotype** and are **heritable** (conserved after cell division).

DNA Methylation

DNA methylation is one of the most studied epigenetic changes.

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



CH_3 CH_3 CH_3

5′ – CAGTACGTCAGCAATCTCGTCGTTACAGCCGTAC – 3′

3′ – GTCATGCAGTCGTTAGAGCAGCAATGTCGGCATG – 5′

CH_3 CH_3 CH_3

DNA Methylation

- methylation patterns can be maintained after cell division.

DNA Methylation

- methylation patterns can be maintained after cell division.
- it depends on the tissue, development stage.

DNA Methylation

- methylation patterns can be maintained after cell division.
- it depends on the tissue, development stage.
- it changes with age and other environmental factors (cancer).

DNA Methylation

- methylation patterns can be maintained after cell division.
- it depends on the tissue, development stage.
- it changes with age and other environmental factors (cancer).
- it can influence gene expression

CpG sites and islands

- CpGs are rare as dinucleotides.

CpG sites and islands

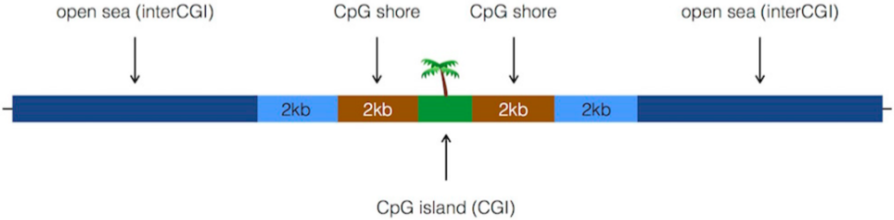
- CpGs are rare as dinucleotides.
- CpGs often cluster in regions of hundreds of base pairs with high CpG concentration, called **CpG islands**.

CpG sites and islands

- CpGs are rare as dinucleotides.
- CpGs often cluster in regions of hundreds of base pairs with high CpG concentration, called **CpG islands**.
- Most CpG islands are unmethylated.

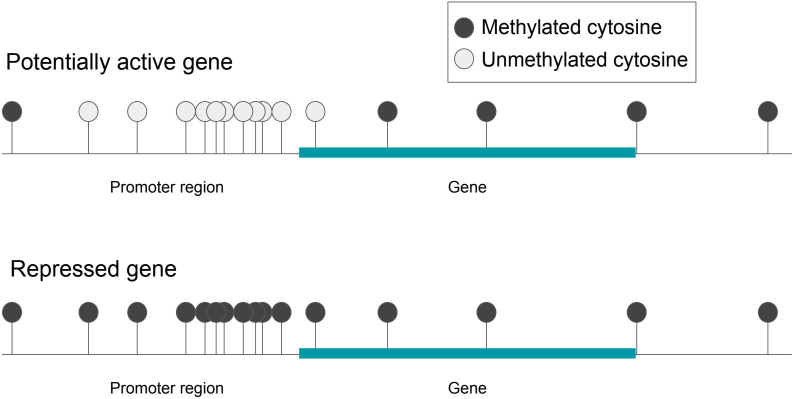
CpG sites and islands

- CpGs are rare as dinucleotides.
- CpGs often cluster in regions of hundreds of base pairs with high CpG concentration, called **CpG islands**.
- Most CpG islands are unmethylated.



DNA methylation regulatory role

promoter methylation leads to gene silencing:



Anyway, CpG islands associated with gene promoters are rarely methylated.

DNA Methylation Role

- outside this specific case, poor understanding of how methylation patterns influence gene expression.

DNA Methylation Role

- outside this specific case, poor understanding of how methylation patterns influence gene expression.
- The most common approach in analyzing methylation data is to measure the mean methylation level for a certain region, and eventually draw conclusions from the observed difference in a sample.

DNA Methylation Role

- outside this specific case, poor understanding of how methylation patterns influence gene expression.
- The most common approach in analyzing methylation data is to measure the mean methylation level for a certain region, and eventually draw conclusions from the observed difference in a sample.
- Genome-wide studies adopting this approach led to a poor correlation with gene expression.

Recent research

Recent research is focusing on the spatial patterns of methylation:

- "sharp" methylation shapes should favor gene expression (Kapourani and Sanguinetti 2016; Jeong et al. 2014; Edgar et al. 2014).
- Positive correlation between expression and gene body methylation (past promoter's island).
- Shores methylation.

Problem Statement

How much does methylation influence gene expression?

Problem Statement

How much does methylation influence gene expression?

Does methylation patterns encode useful information?

Problem Statement

How much does methylation influence gene expression?

Does methylation patterns encode useful information?

We explored the application of *multiscale relevance*, a recently developed information-theoretic method.

Table of Contents

1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

Overview

MSR (*Multi-Scale Relevance*) is a function that associates to a set of M real numbers a real number: $\mathbb{R}^M \to \mathbb{R}$

Overview

MSR (*Multi-Scale Relevance*) is a function that associates to a set of M real numbers a real number: $\mathbb{R}^M \to \mathbb{R}$

- intuitively, it gives a measure of the richness of density states at different scales of a set of real values.

Overview

MSR (*Multi-Scale Relevance*) is a function that associates to a set of M real numbers a real number: $\mathbb{R}^M \to \mathbb{R}$

- intuitively, it gives a measure of the richness of density states at different scales of a set of real values.
- first defined in Cubero, Marsili, et al. 2020 for identifying informative neurons.

Overview

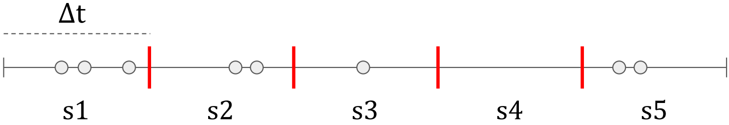
MSR (*Multi-Scale Relevance*) is a function that associates to a set of M real numbers a real number: $\mathbb{R}^M \to \mathbb{R}$

- intuitively, it gives a measure of the richness of density states at different scales of a set of real values.
- first defined in Cubero, Marsili, et al. 2020 for identifying informative neurons.
- motivation rooted in a series of articles on criticality of efficient representations (Marsili et al. 2013; Haimovici and Marsili 2015; Cubero, Jo, et al. 2019).

Definition

Given the set of M real values $\mathbf{t}$, we can define a compressed representation $\mathbf{s} = s(\mathbf{t})$ based on a subdivision in bins of size Δt.

Example:



t	4.8	5.4	6.5	9.1	9.6	12.3	18.5	19.0
s(**t**)	s1	s1	s1	s2	s2	s3	s5	s5

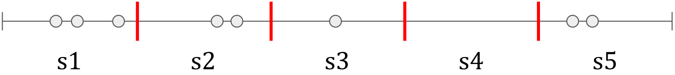
Resolution

Instead of Δt we define as *resolution* the entropy of $\mathbf{s}$:

$$H[s] = -\sum_{s=1}^{T} \frac{k_s}{M} log_M \frac{k_s}{M}$$

where k_s is the number of values inside the bin s.

Example:



t	4.8	5.4	6.5	9.1	9.6	12.3	18.5	19.0
s	s1	s1	s1	s2	s2	s3	s5	s5

$\rightarrow H[s]$

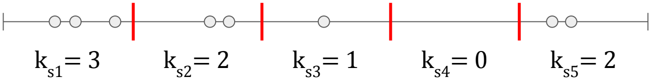
Relevance

We define as *relevance*:

$$H[K] = -\sum_{k=1}^{M} \frac{km_k}{M} log_M \frac{km_k}{M}$$

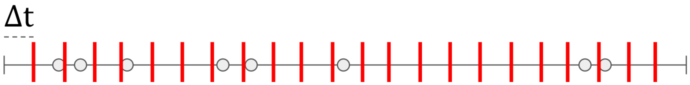
where m_k indicates the number of bins containing k values.

Example:

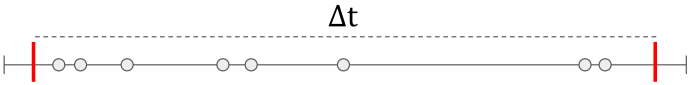


t	4.8	5.4	6.5	9.1	9.6	12.3	18.5	19.0	
s	s1	s1	s1	s2	s2	s3	s5	s5	→ *H[s]*
$\mathbf{k_s}$	3	3	3	2	2	1	2	2	→ *H[k]*

Multiscale relevance



s	s2	s3	s5	s8	s9	s12	s20	s21	→ *H[s]=1*
k	1	1	1	1	1	1	1	1	→ *H[k]=0*



s	s1	s1	s1	s1	s1	s1	s1	s1	→ *H[s]=0*
k	8	8	8	8	8	8	8	8	→ *H[k]=0*

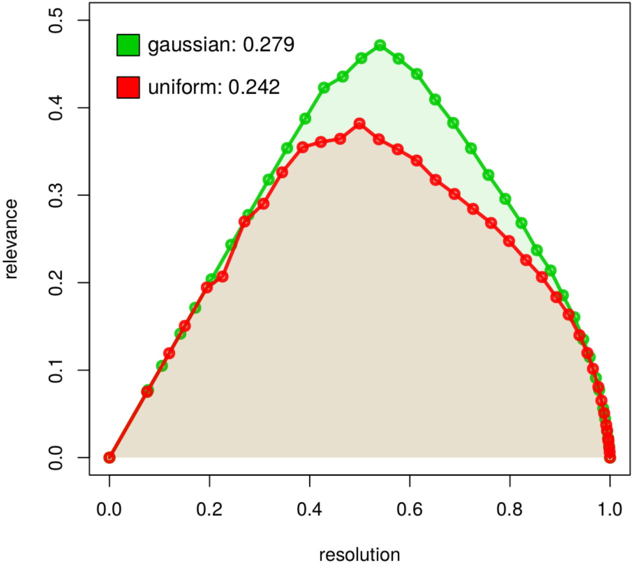
Multiscale relevance

- For intermediate values of Δt, $H[s]$ spans from 0 to 1 and $H[K]$ assumes positives values.

Multiscale relevance

- For intermediate values of Δt, $H[s]$ spans from 0 to 1 and $H[K]$ assumes positives values.
- As we vary Δt, we can trace a curve in the $H[s] - H[K]$ space and calculate the area under the curve, that we call *Multi-Scale Relevance* (MSR).

Example


🟩 gaussian: 0.279
🟥 uniform: 0.242
relevance
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
resolution

Theoretical motivation

It has been shown that the *relevance* $H[k]$ has several properties, for example:

Theoretical motivation

It has been shown that the *relevance* $H[k]$ has several properties, for example:

- $H[k]$ provides an upper bound to the information the data contains on the generative process (Cubero, Jo, et al. 2019).

Theoretical motivation

It has been shown that the *relevance* $H[k]$ has several properties, for example:

- $H[k]$ provides an upper bound to the information the data contains on the generative process (Cubero, Jo, et al. 2019).
- broad distributions emerge when $H[k]$ is maximized at fixed $H[s]$.

Theoretical motivation

It has been shown that the *relevance* $H[k]$ has several properties, for example:

- $H[k]$ provides an upper bound to the information the data contains on the generative process (Cubero, Jo, et al. 2019).
- broad distributions emerge when $H[k]$ is maximized at fixed $H[s]$.

So MSR provides a summary of $H[k]$ at multiple scales.

Why MSR on methylation data?

Why MSR on methylation data?

- The suggestion is that methylation could encode information through its spatial patterns. Then MSR would measure the ”information” relative to a certain region.

MSR characteristics

- It can be proved that the maximum value for MSR is ≈ 0.3 .
- MSR is too noisy when $M < 100$, it requires large samples.

Table of Contents

1 Genetics Recap

2 Epigenetics and DNA Methylation

3 MSR

4 applying MSR to methylation data

5 Genomewide MSR application

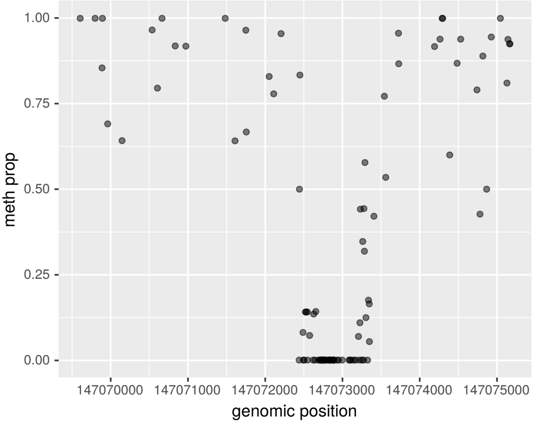
6 Relationship with expression

7 Conclusions

Methylation Data

Whole Genome Bisulfite Sequencing (WGBS) provides genome-wide methylation information at single CpG resolution.

	chr	pos	strand	reads	prop
1	chr1	..798	+	16	1.00
2	chr1	..799	-	5	1.00
3	chr1	..888	+	6	0.83
4	chr1	..889	-	1	1.00
5	chr1	..893	+	6	1.00
6	chr1	..894	-	1	1.00
7	chr1	..961	+	7	1.00
8	chr1	..962	-	6	0.33

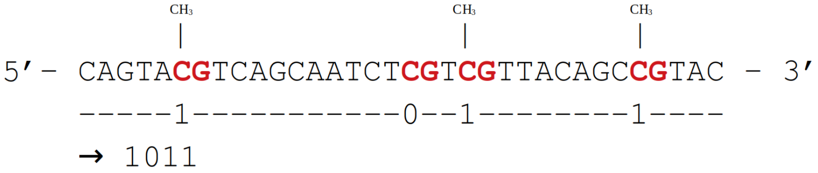


Methylation Data Representation

Methylation can be represented as a binary string:

Methylation Data Representation

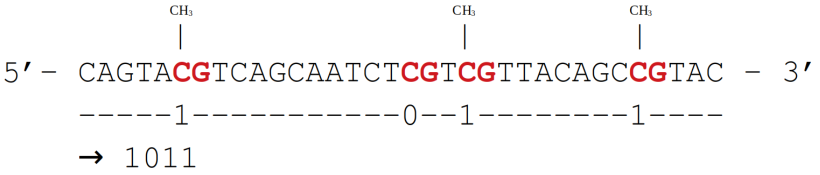
Methylation can be represented as a binary string:
CpG list: methylated vs unmethylated CpG



```
          CH3                            CH3               CH3
           |                              |                 |
5'- CAGTACGTCAGCAATCTCGTCGTTACAGCCGTAC - 3'
    -----1-----------0--1--------1----
    → 1011
```

Methylation Data Representation

Methylation can be represented as a binary string:
CpG list: methylated vs unmethylated CpG



```
5′ – CAGTACGTCAGCAATCTCGTCGTTACAGCCGTAC – 3′
     -----1-----------0--1--------1----
     → 1011
```

- MSR_1 = MSR on indexes of methylated CpGs
- MSR_0 = MSR on indexes of unmethylated CpGs

Binarization

- strand information is ignored
- proportion is binarized (0.5 threshold)

pos	strand	reads	prop
78	+	10	0.2
79	-	12	0.25
107	+	2	1
108	-	0	-
130	+	4	1
131	-	6	0.5
132	+	4	0.75
133	-	0	-

→

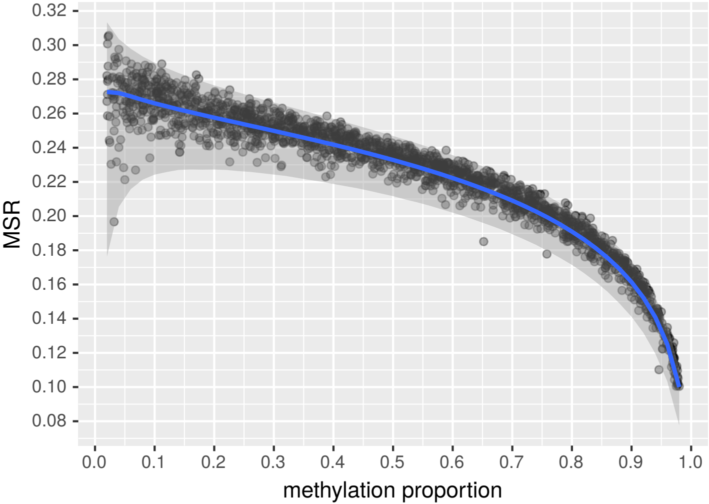
pos	reads	prop
78	22	0.27
107	2	1
130	10	0.7
132	4	0.75

→

pos	state
78	0
107	1
130	1
132	1

MSR with discrete positions

MSR for random binary strings of length 1000 of various mean "methylation" level.



MSR with discrete positions

- ”saturation” for high densities of ones causes low MSR.

MSR with discrete positions

- "saturation" for high densities of ones causes low MSR.
- we derived statistics from *MSR*, in order "adjust" it according to the proportion of methylated sites.

MSR with discrete positions

- "saturation" for high densities of ones causes low MSR.
- we derived statistics from *MSR*, in order "adjust" it according to the proportion of methylated sites.
- *residual* := difference between MSR and median value for that density of ones.

MSR with discrete positions

- "saturation" for high densities of ones causes low MSR.
- we derived statistics from *MSR*, in order "adjust" it according to the proportion of methylated sites.
- *residual* := difference between MSR and median value for that density of ones.
- *ecdf* := probability to obtain randomly a value smaller than the observed one (for that density).

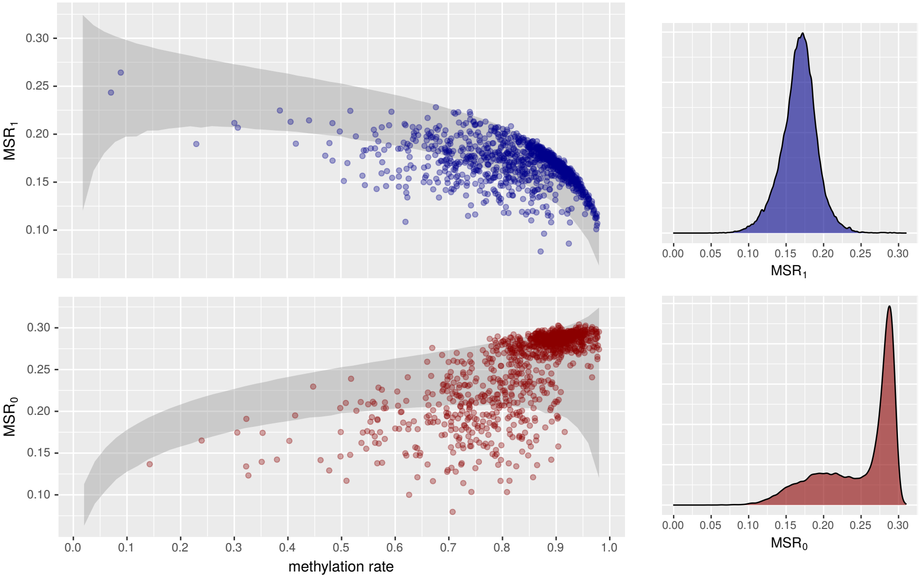
Table of Contents

1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

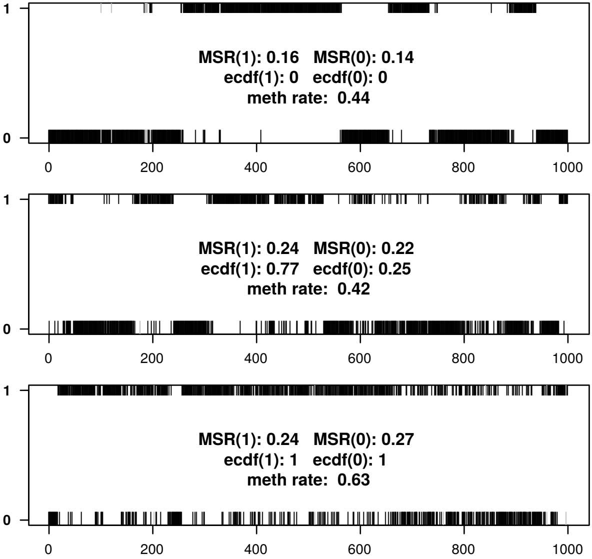
Genomewide MSR application

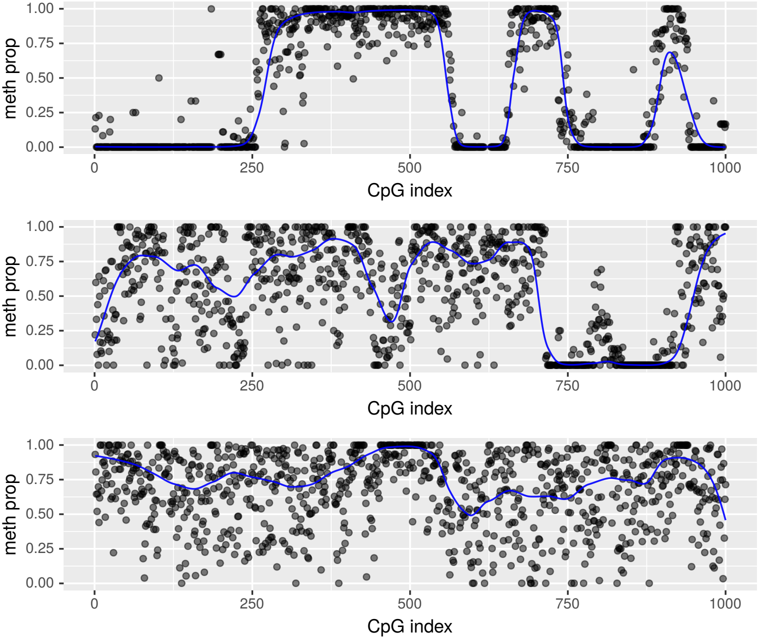
We divide the genome (as example we used stomach tissue WGBS from ENCODE) in fragments of 1000 CpGs and then we calculated MSR.

MSR distribution

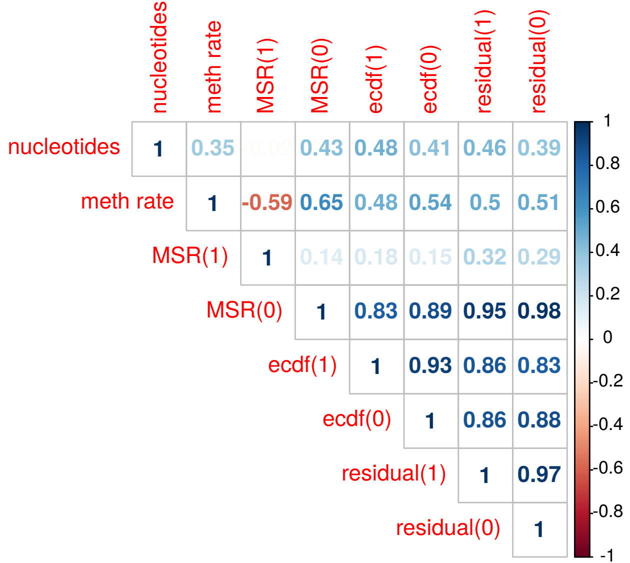


Visual intuition




meth prop
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
CpG index

Correlation between fragments features (Pearson's r):



	nucleotides	meth rate	MSR(1)	MSR(0)	ecdf(1)	ecdf(0)	residual(1)	residual(0)
nucleotides	1	0.35	-0.02	0.43	0.48	0.41	0.46	0.39
meth rate		1	-0.59	0.65	0.48	0.54	0.5	0.51
MSR(1)			1	0.14	0.18	0.15	0.32	0.29
MSR(0)				1	0.83	0.89	0.95	0.98
ecdf(1)					1	0.93	0.86	0.83
ecdf(0)						1	0.86	0.88
residual(1)							1	0.97
residual(0)								1

Methylation autocorrelation

- MSR_0 and the related statistics measures regularity of methylation patterns.

Methylation autocorrelation

- MSR_0 and the related statistics measures regularity of methylation patterns.
- In this context regularity seems to coincide with the similarity of contiguous CpGs.

Methylation autocorrelation

- MSR_0 and the related statistics measures regularity of methylation patterns.
- In this context regularity seems to coincide with the similarity of contiguous CpGs.
- auto-correlation of contiguous CpGs' methylation proportion should capture this characteristic.

Methylation autocorrelation

Methylation autocorrelation is highly correlated with several MSR features (Pearson's r):

	nucleotides	meth rate	CpG sites MSR	MSR(1)	MSR(0)	ecdf(1)	ecdf(0)	residual(1)	residual(0)
meth autocorrelation	-0.59	-0.56	0.36	-0.05	-0.84	-0.89	-0.85	-0.86	-0.82

Table of Contents

1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

Relationship with expression

- Objective: predict transcriptional activity relative to a certain region.

Relationship with expression

- Objective: predict transcriptional activity relative to a certain region.
- We need to assign to each fragment a measure of its transcriptional activity.

Relationship with expression

- Objective: predict transcriptional activity relative to a certain region.
- We need to assign to each fragment a measure of its transcriptional activity.
- Same experiments for several cell types: H1, endodermal, K562, GM12878, GM23248, HeLa, lung, stomach.

Gene expression data

- Rna-seq provides measures of transcriptional activity for each gene.

Gene expression data

- Rna-seq provides measures of transcriptional activity for each gene.
- We mainly use polyA plus Rna-seq data, that focuses on the set of protein coding genes ($\approx 20,000$).

Gene expression data

- Rna-seq provides measures of transcriptional activity for each gene.
- We mainly use polyA plus Rna-seq data, that focuses on the set of protein coding genes ($\approx 20,000$).
- TPM (Transcript Per Million) measures the relative abundance of RNAs. In particular we use $\log_2(\text{TPM}+\epsilon)$.

Gene expression data

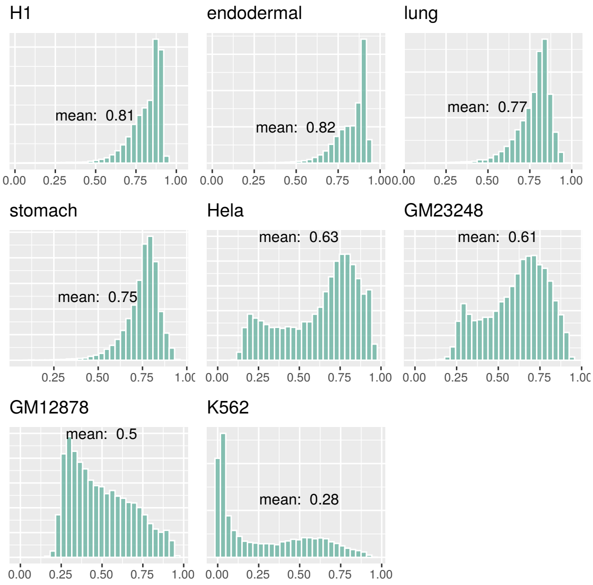
- Rna-seq provides measures of transcriptional activity for each gene.
- We mainly use polyA plus Rna-seq data, that focuses on the set of protein coding genes ($\approx 20,000$).
- TPM (Transcript Per Million) measures the relative abundance of RNAs. In particular we use $\log_2(\text{TPM}+\epsilon)$.
- We assign to each fragment the sum of TPMs of genes having their transcription start site in that region.

Features

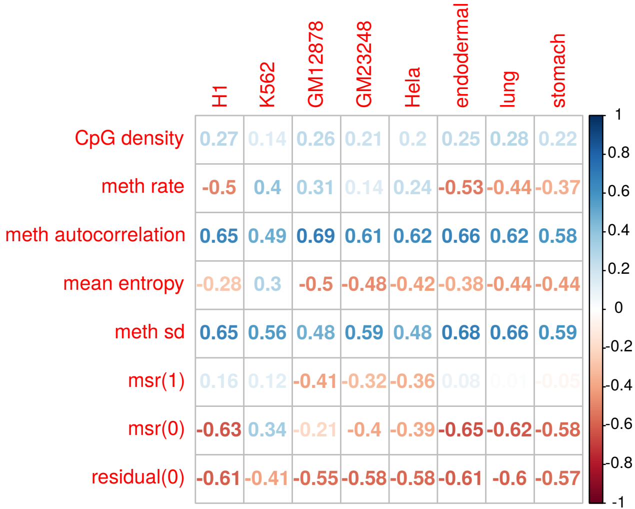
We divide features in three groups:

- **Basic**: mean methylation level, nucleotides, CpG density.
- **Advanced**: methylation autocorrelation, methylation mean entropy, methylation standard deviation.
- **MSR related**

Overall Methylation levels

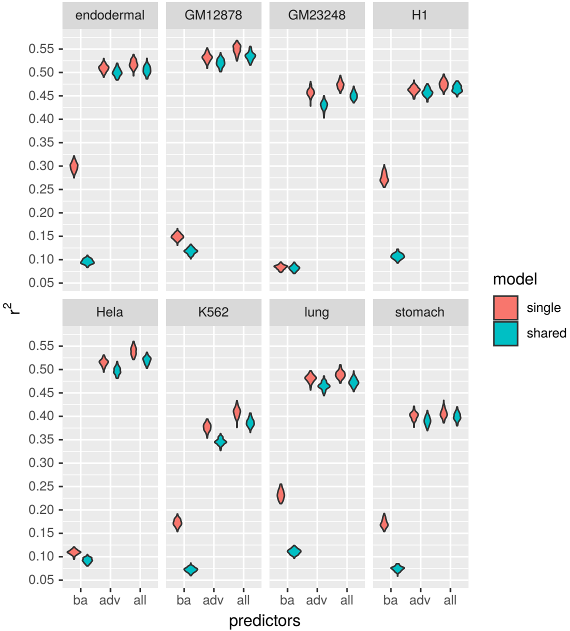


Pearson's r between features and expression for different cells



	H1	K562	GM12878	GM23248	Hela	endodermal	lung	stomach
CpG density	0.27	0.14	0.26	0.21	0.2	0.25	0.28	0.22
meth rate	-0.5	0.4	0.31	0.14	0.24	-0.53	-0.44	-0.37
meth autocorrelation	0.65	0.49	0.69	0.61	0.62	0.66	0.62	0.58
mean entropy	-0.28	0.3	-0.5	-0.48	-0.42	-0.38	-0.44	-0.44
meth sd	0.65	0.56	0.48	0.59	0.48	0.68	0.66	0.59
msr(1)	0.16	0.12	-0.41	-0.32	-0.36	0.08	0.01	-0.05
msr(0)	-0.63	0.34	-0.21	-0.4	-0.39	-0.65	-0.62	-0.58
residual(0)	-0.61	-0.41	-0.55	-0.58	-0.58	-0.61	-0.6	-0.57

- meth autocorrelation is in general the most correlated.
- meth rate correlation sign depends on tissue



Test R^2 (several splits) for linear models with different sets of predictors

- **single models** are fitted on a single cell type datasets
- **shared model** is fitted on a dataset including all cells, and then evaluated separately for each cell type

Discussion

- models with only basic predictors have poor performances and don't generalize.

Discussion

- models with only basic predictors have poor performances and don't generalize.
- adding meth. autocorrelation and meth. sd let models explain almost half of the variance for several cells.

Discussion

- models with only basic predictors have poor performances and don't generalize.
- adding meth. autocorrelation and meth. sd let models explain almost half of the variance for several cells.
- MSR features add little information about expression.

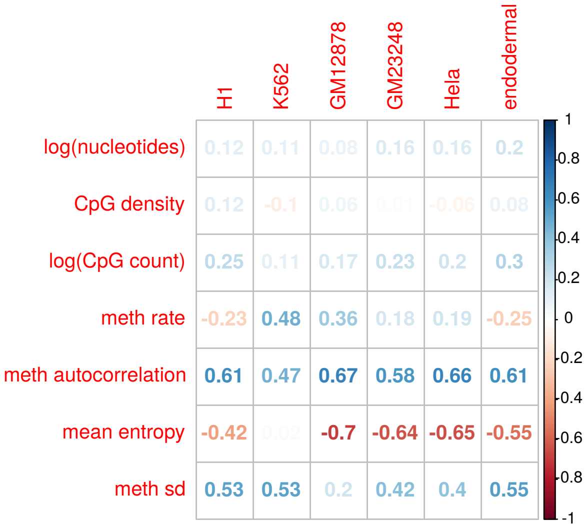
Relationship at Gene level

Now we repeat the same experiment but focusing on genes:

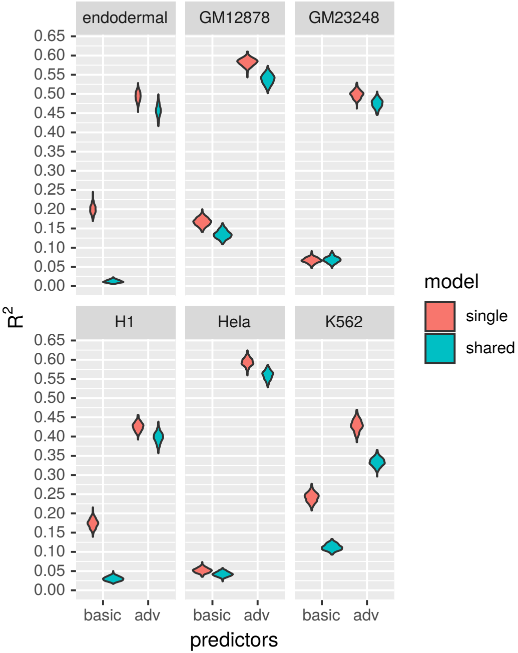
- Focus on gene bodies methylation.
- Only cell lines are considered.
- This time we don't consider MSR, since the number of CpGs in gene bodies is variable, and often too small.

Features correlation with expression

Pearson's *r* between features and expression for different cells



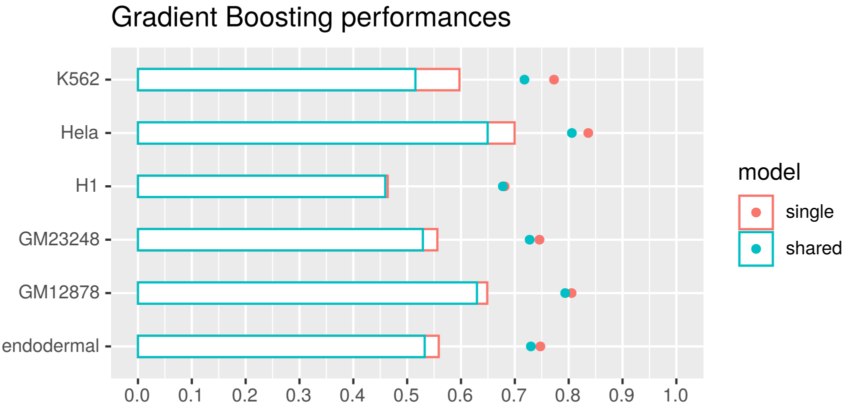
	H1	K562	GM12878	GM23248	Hela	endodermal
log(nucleotides)	0.12	0.11	0.08	0.16	0.16	0.2
CpG density	0.12	-0.1	0.06	0.01	-0.06	0.08
log(CpG count)	0.25	0.11	0.17	0.23	0.2	0.3
meth rate	-0.23	0.48	0.36	0.18	0.19	-0.25
meth autocorrelation	0.61	0.47	0.67	0.58	0.66	0.61
mean entropy	-0.42	0.02	-0.7	-0.64	-0.65	-0.55
meth sd	0.53	0.53	0.2	0.42	0.4	0.55



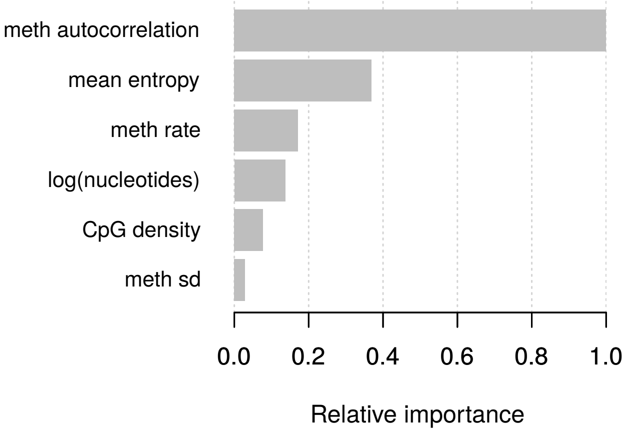
Test R^2 (several splits) for linear models with different sets of predictors

- **single models** are fitted on different cells types datasets
- **shared model** is fitted on a dataset including all cells, and then evaluated separately for each cell type

Gradient Boosting



Performances of a tree-based model fitted with Gradient Boosting (Bars are test R^2, points are Pearson's r)



The relative importance in gradient boosting is based on the number of times a variable is selected for splitting, and on the improvement to the model as a result of each split (Elith et al. 2008).

Discussion

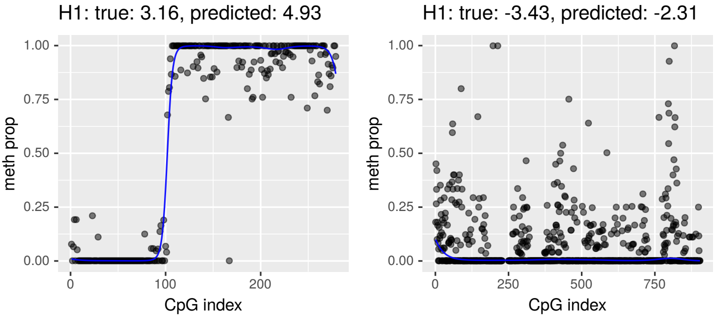
- Significant improvement with respect to trivial models.

Discussion

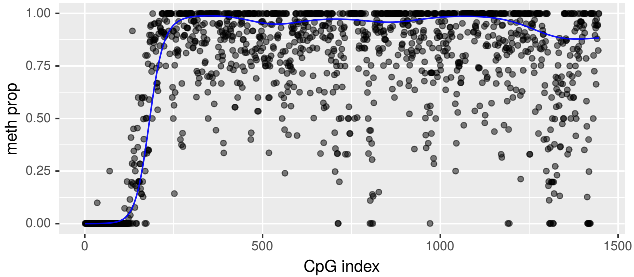
- Significant improvement with respect to trivial models.
- Gene expression is generally higher where there is:
  - correlation between methylation state of contiguous CpGs.
  - homogeneity in methylation between cells.
  - not too low or too high overall methylation level.

Discussion

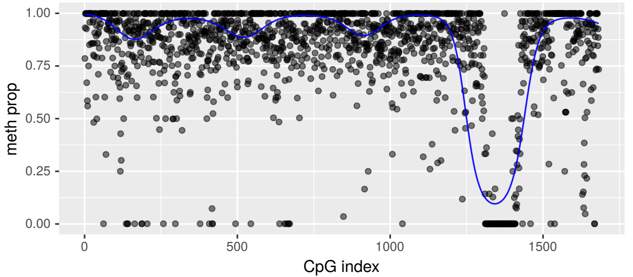
- Significant improvement with respect to trivial models.
- Gene expression is generally higher where there is:
    - correlation between methylation state of contiguous CpGs.
    - homogeneity in methylation between cells.
    - not too low or too high overall methylation level.



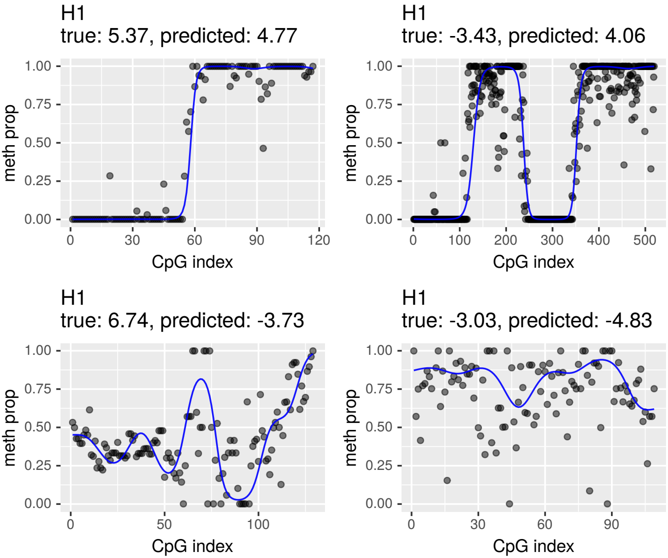
H1: true: 1.01, predicted: 1.71



H1: true: 0.67, predicted: 1.24



There are still several regions with a misleading methylation pattern according to our models.



H1
true: 5.37, predicted: 4.77

H1
true: -3.43, predicted: 4.06

H1
true: 6.74, predicted: -3.73

H1
true: -3.03, predicted: -4.83

Table of Contents

1. Genetics Recap
2. Epigenetics and DNA Methylation
3. MSR
4. applying MSR to methylation data
5. Genomewide MSR application
6. Relationship with expression
7. Conclusions

Results

- We found methylation characteristics that are determinant in gene expression.

Results

- We found methylation characteristics that are determinant in gene expression.
- Dramatic improvement with respect to the model that only consider mean methylation level and CpG density.

Results

- We found methylation characteristics that are determinant in gene expression.
- Dramatic improvement with respect to the model that only consider mean methylation level and CpG density.
- Models hold for arbitrary regions.

Results

- We found methylation characteristics that are determinant in gene expression.
- Dramatic improvement with respect to the model that only consider mean methylation level and CpG density.
- Models hold for arbitrary regions.
- These findings are coherent with recent research.

Limits

- Focus on the correct functional region could be needed.

Limits

- Focus on the correct functional region could be needed.
- Good performances in some cancer cells suggest that those features may be useful in detecting "degenerated" genes more than differences between "healthy" genes.

Limits

- Focus on the correct functional region could be needed.
- Good performances in some cancer cells suggest that those features may be useful in detecting "degenerated" genes more than differences between "healthy" genes.
- The difference in expression of different genes in a cell is due also to genomic features.

Limits

- Focus on the correct functional region could be needed.
- Good performances in some cancer cells suggest that those features may be useful in detecting ”degenerated” genes more than differences between ”healthy” genes.
- The difference in expression of different genes in a cell is due also to genomic features.
- We focused mainly on the relative positions of methylated and unmethylated sites, ignoring their spatial distribution.

Future directions

- Probably further improvements are possible if considering larger areas around genes.

Future directions

- Probably further improvements are possible if considering larger areas around genes.
- There could be useful information in the size, the number, and the location of islands.

Future directions

- Probably further improvements are possible if considering larger areas around genes.
- There could be useful information in the size, the number, and the location of islands.
- Investigate the role of these features in tissue specific genes.

Future directions

- Probably further improvements are possible if considering larger areas around genes.
- There could be useful information in the size, the number, and the location of islands.
- Investigate the role of these features in tissue specific genes.
- Verify these results comparing the same gene in a population.

What about MSR?

- MSR resulted useful in detecting hyper-regular structures.

What about MSR?

- MSR resulted useful in detecting hyper-regular structures.
- MSR on CpG list was often well correlated with expression, but it was not more useful than other a posteriori extracted features in predicting expression.

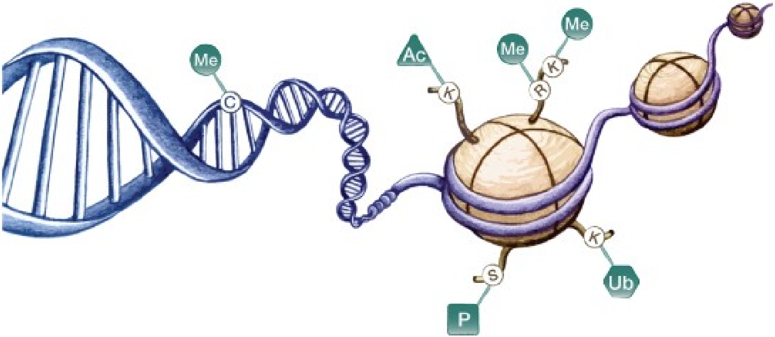
What about MSR?

- MSR resulted useful in detecting hyper-regular structures.
- MSR on CpG list was often well correlated with expression, but it was not more useful than other a posteriori extracted features in predicting expression.
- MSR applied on genomic positions of methylated or unmethylated sites has still to be explored.

What about MSR?

- MSR resulted useful in detecting hyper-regular structures.
- MSR on CpG list was often well correlated with expression, but it was not more useful than other a posteriori extracted features in predicting expression.
- MSR applied on genomic positions of methylated or unmethylated sites has still to be explored.
- MSR could be related to other covariates.

Thank you!



References I

- Ryan John Cubero, Junghyo Jo, Matteo Marsili, Yasser Roudi, and Juyong Song. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.6 (2019), p. 063402.

- Ryan John Cubero, Matteo Marsili, and Yasser Roudi. In: *Journal of computational neuroscience* 48.1 (2020), pp. 85–102.

- Rachel Edgar, Powell Patrick Cheng Tan, Elodie Portales-Casamar, and Paul Pavlidis. In: *Epigenetics & chromatin* 7.1 (2014), p. 28.

References II

Jane Elith, John R Leathwick, and Trevor Hastie. In: *Journal of Animal Ecology* 77.4 (2008), pp. 802–813.

european-biotechnology.com. https://european-biotechnology.com/up-to-date/latest-news/news/epigenetic-drugs-set-to-boost-immunoncology.html. 2017.

Silvia Grigolon, Silvio Franz, and Matteo Marsili. In: *Molecular BioSystems* 12.7 (2016), pp. 2147–2158.

Ariel Haimovici and Matteo Marsili. In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.10 (2015), P10013.

References III

- Mira Jeong, Deqiang Sun, Min Luo, Yun Huang, Grant A Challen, Benjamin Rodriguez, Xiaotian Zhang, Lukas Chavez, Hui Wang, Rebecca Hannah, et al. In: *Nature genetics* 46.1 (2014), pp. 17–23.

- ChantrioInt-Andreas Kapourani and Guido Sanguinetti. In: *Bioinformatics* 32.17 (2016), pp. i405–i412.

- Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.09 (2013), P09003.

- Juyong Song, Matteo Marsili, and Junghyo Jo. In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.12 (2018), p. 123406.