UNIVERSITÀ DEGLI STUDI DI TRIESTE

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Laurea Magistrale in Data Science and Scientific Computing

Tesi di Laurea Magistrale

AN INFORMATION-THEORETIC INVESTIGATION INTO EPIGENETIC REGULATION OF GENE EXPRESSION

Laureando:
Davide Scassola

Relatore:
Guido Sanguinetti

Correlatore:
Matteo Marsili

ANNO ACCADEMICO 2019–2020

Abstract

L'epigenetica è lo studio dei cambiamenti ereditabili da una cellula che influenzano il fenotipo di un individuo, senza modificarne la sequenza di DNA. In particolare la metilazione del DNA, che consiste nella modificazione di una base azotata, è uno dei tratti epigenetici più studiati. Tuttavia si conosce ancora poco su come la metilazione influenzi l'espressione genica. Le analisi si sono focalizzate principalmente sulla differenza nei livelli medi di metilazione di specifiche zone, mentre recenti studi hanno sottolineato l'importanza della distribuzione spaziale dei siti metilati. In questa tesi si esplora l'utilizzo di un indicatore recentemente introdotto (MSR) basato sulla teoria dell'informazione, con l'idea che la disposizione nello spazio dei siti metilati possa codificare informazione utile. Le motivazioni si fondano su una recente serie di articoli che indaga l'emergere di determinate proprietà statistiche in sistemi che ottimizzano l'informazione. L'applicazione di MSR ad un intero metiloma diviso per zone fa emergere alcune caratteristiche, principalmente la correlazione nei livelli di metilazione dei siti contigui, che si dimostra positivamente correlata con l'espressione. Alla luce di queste considerazioni si costruiscono dei modelli "genome-wide" per la predizione dell'espressione genica che migliorano significativamente l'accuratezza dei modelli basati sul solo livello medio di metilazione. I risultati ottenuti suggeriscono che l'attività trascrizionale è generalmente favorita nelle zone in cui c'è compresenza di aree metilate e non metilate, la separazione tra queste è netta e la metilazione è coerente tra le cellule del campione.

Table of contents

Introduction

Each complex organism is made of a multitude of cells, each one having a copy of the same genetic information. This information is stored in a molecule named DNA, in the form of a long sequence of 4 possible bases. Gene expression is the translation of this information into specific molecules, RNAs and the corresponding proteins, functional to the proper living of the organism. RNAs that are produced by cells of a certain organism can vary significantly from cell to cell, especially those with a different function, in addition more than 95% of DNA information is never translated into RNA. Although all cells shares the same genetic information, it seems each cell is able to select and use the correct information contained in its DNA, according to its specific role in the organism.

This is made possible by molecular mechanisms that involve DNA without altering its base sequence, for example making some regions inaccessible to enzymes involved in translation. Epigenetics studies these kind of heritable phenotype changes and it's of fundamental importance in our understanding of the gene expression machinery. One of the most studied is DNA methylation, that consists in the addition of a molecular group to a base, and it's associated with several relevant biological processes.

There is a well documented association with hyper-methylation of functional regions and transcriptional repression. However, outside this specific case, the understanding of how methylation patterns influence gene expression is considerably weaker. The most common approach in analyzing methylation data is to measure the difference in the mean methylation level for a certain region. Yet genome-wide studies adopting this approach often led to a poor correlation with gene expression. In fact more recent studies are trying to exploit the availability of high-resolution data (the knowledge of methylation state at base resolution) to investigate a possible relation of spatial patterns of methylated sites with expression. A more precise investigation of the quantitative relation between methylation patterns and expression could deepen our understanding on the functioning of this epigenetic regulation mechanism and quantify its relative importance.

In this thesis the application on methylation data of a recently developed information-theoretic method is explored, driven by the idea that methylation could encode useful information through the spatial distribution of methylated sites. This method is inspired by a number of recent publications that investigated the emergence of statistical criticality,

i.e. the occurrence of power law frequency distributions, in systems that extract efficient representations.

In chapter 1 a general introduction to genetics, epigenetics and DNA methylation is provided. In particular it focuses on DNA methylation characteristics and its known functions and mechanisms.

In chapter 2 the *multiscale relevance* (MSR) is defined, followed by the information-theoretic arguments that led to its introduction.

Chapter 3 introduces methylation data and discusses how it's manipulated in order to apply MSR. Then, after some preliminary considerations, MSR is applied to a methylation dataset as a first exploratory analysis.

Finally in chapter 4, starting from the considerations of the previous chapter, a quantitative relation between transcriptional activity and MSR with other methylation features is studied. The aim is to evaluate both the importance of different methylation characteristics and the goodness of models based on these in predicting expression.

Chapter 1

DNA Methylation

In the first sections of this chapter an introduction to genetics is given (based on [2, 22, 1]). Then we proceed with sections about DNA methylation, which are mainly based on material on the book by Tost [21] and on the review article Moore et al. [17].

1.1 DNA

Our bodies are made up of trillions of cells. Each cell, with few exceptions, contains in its nucleus a complete copy of an individual's genetic information, also known as genome. This information is fundamental to define every biological process in the organism and determine the phenotype, i.e the set of all its observable characteristics or traits.

The genetic information is stored in a molecule named DNA (DeoxyriboNucleic Acid), that is composed by a sequence, or chain, of paired nucleotides. In the DNA there are 4 type of nucleotides. What differentiate them is a component named nitrogenous base that can be of four types: adenine (A), thymine (T), guanine (G) and cytosine (C). So genetic information is encoded in the particular sequence of nucleotides, equivalently, in the particular sequence of bases, as the sequence of letters of a text written in a language based on a 4 letters alphabet.

DNA is not composed by a single chain of nucleotides. Each nucleotide is paired according to a precise rule: the complementary base of adenine is thymine, and the one of guanine is cytosine, giving rise in this way to another chain of nucleotides, fully determined by the other one. This permits the DNA to duplicate since the "complementary" chain can be exactly reconstructed from the other.

In the DNA the two complementary chains are also referred as "strands", and by convention, one of them is called "plus strand" and the other "minus strand", in order to have an unambiguous way to refer to a genomic sequence in the DNA. [1]

[1] A more used way to refer to the "plus" direction is the notation 5'$\rightarrow$ 3'

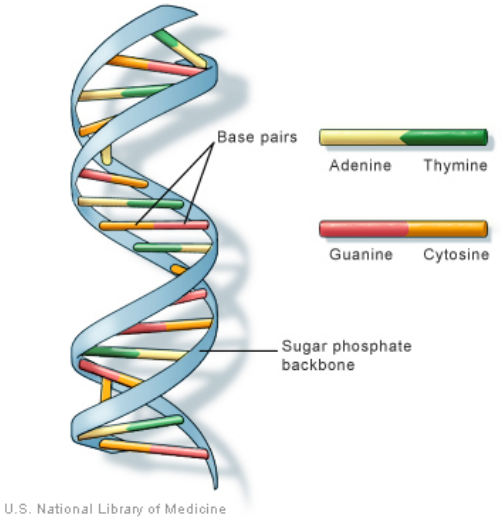

Fig. 1.1 ([15]) The shape of the DNA molecule resemble a double helix.

In general with "nucleic acids" are referred both paired sequences of nucleotides (DNA) and single chains of nucleotides (RNA). A difference in RNA is that the uracil, another nitrogenous base, takes the place of the thymine.

All genetic information inside the cell nucleus is not placed in a single DNA molecule, instead it is subdivided in chromosomes. Chromosomes are very long strands of DNA, coiled up and packaged in such a way as to occupy a relatively small space. In humans there are 46 chromosomes contained in the nucleus of each body cells, in particular they are 23 pairs of (pairwise) similar chromosomes, in each pair one of the two chromosomes comes from the mother and the other from the father. In humans one of these 23 pairs are called sex chromosomes, since they exists in two variants that determine the sex of the individual. These two variants are referred as XY chromosomes for males, and XX for females. The genome intended as the whole information contained in the 46 chromosomes is referred as diploid genome, instead if it is considered only one of the pairs (23 chromosomes) it is referred as haploid genome.

The way the genome influences the life of an organism is through the synthesis of proteins, large biomolecules composed by a long chain of amino acid residues that perform a vast array of functions. Genomic information can be translated into proteins thanks to the genetic code, a mapping from the set of possible triplets of bases to the set of all the amino acids, this allows to translate a sequence of bases into the sequence of amino acids that will form a protein.

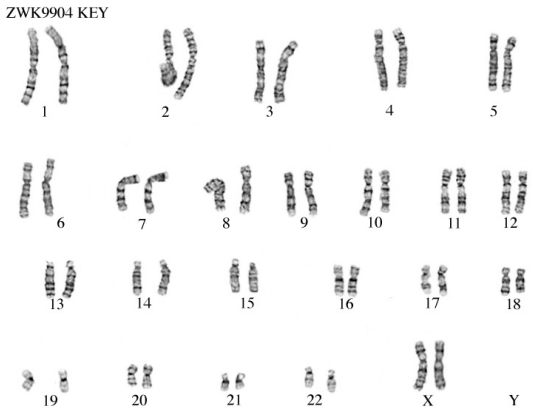

Fig. 1.2 ([20]) *The 46 chromosomes of a female individual.*

1.2 Gene Expression

Though all the genome is made up of billions of nucleotides, about the 98% does not encode protein sequences. The rest is found in functional segments of DNA named "genes". In the human genome there are between 20.000 and 25.0000 genes each one having its own specific location on a chromosome. Despite the fact that intra-genic DNA is often called "silent" or "junk" DNA, there is evidence that it can have different functions, such as regulating gene expression.

Gene expression is the process by which the information contained in a gene is used for the synthesis of functional molecules. These products are often proteins, but there exists also non-protein-coding genes that produce for example functional RNA molecules also named non-coding RNAs (ncRNAs). The first of several steps of gene expression of proteins is transcription, during which a particular segment of DNA is copied into a molecule of RNA named messenger-RNA (mRNA). mRNA is then translated into a protein.

The central dogma of molecular biology states that "information" can only be passed from nucleic acid to nucleic acid, or from nucleic acid to protein, but not from proteins to nucleic acids. Where information means the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein (Crick [5]). So genomic information can only be contained in nucleic acids.

Gene expression is the most fundamental level at which the genotype gives rise to the phenotype, and characterizes all known organism. The information the genome contains represents the genotype, whereas the phenotype results from the "interpretation" of that information. The phenotype is often expressed by the synthesis of proteins that control the organism's structure and development, or that act as enzymes catalyzing specific reactions.

The gene expression process may be regulated, in the sense that the expression a given gene product can be changed in timing, location, and amount, with the possibility to have

a deep effect on the cellular structure and function. Gene expression regulation is at the basis of cellular differentiation and development, and then of versatility and adaptability of any organism.

1.3 Chromatin

Chromatin is a molecular complex that gives shape to a chromosome, and is composed of DNA and other proteins called "histones". Its function is packaging long DNA molecules into compact and dense structures, DNA by itself would be much longer and more fragile. Histones are a family of small proteins termed H1, H2A, H2B, H3, and H4. The basic modular unit of chromatin is the nucleosome. The nucleosome is composed by an octamer (eight proteins) of two each of the histones H2A, H2B, H3, and H4, around which about 146 base pairs of DNA are wrapped (about 1.7 turns). Each chromosome is thus a long chain of nucleosomes, with the appearance of a string of beads, that is further coiled into an even shorter, thicker fiber. Depending on the context/location, the chromatin can be found in an open or lightly packed configuration (euchromatin), or in a closed, tightly packed configuration (heterochromatin).

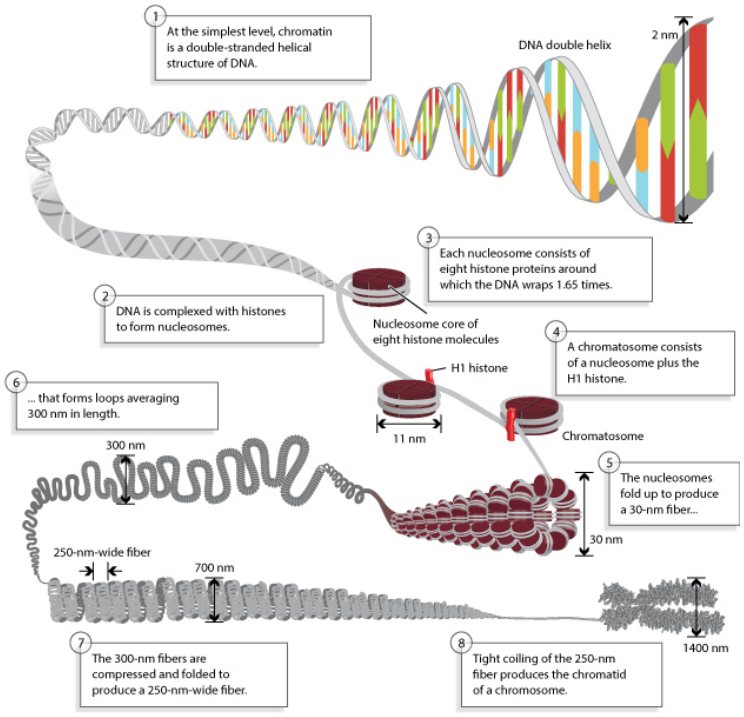

Fig. 1.3 (Annunziato [1]) *The multi-level packaging of chromatin.*

There are some processes such as transcription and replication that require the two strands of DNA to come apart temporarily, it is therefore important for cells to be able to open up chromatin fibers and/or to remove histones transiently. There are two major

processes that make chromatin more accessible: the enzymatic modification of histones through the addition of methyl, acetyl, or phosphate groups, or their displacement by chromatin remodeling complexes. These processes are reversible. Remodeled or modified chromatin can go back to its compact state after transcription or replication are complete.

1.4 Epigenetics

Although in an organism all cells share the same genetic material in terms of DNA base sequence, there is a functional and morphological heterogeneity, essential to the definition of the organism. Epigenetic changes are responsible of this diversity through differential gene expression patterns. Epigenetics is defined as the study of heritable phenotype changes that are not caused by a difference in the primary DNA sequence (the bases sequence), like the gene expression pattern that characterize a specific cell type. These changes are heritable in the sense that are conserved after mitosis (cell duplication), and in some cases also after meiosis (generation of gametes from a body cell). Genomic adaption to an environment is complemented by epigenetic regulation, that represents the link between the genetic code and the phenotype.

Epigenetics consists of several molecular mechanisms that are different but at the same time are closely related and stabilize each other in order to maintain an epigenetic state through time and particularly through cell division. Nevertheless, epigenetic states are not definitive and modifications can occur both as a response to environmental stimuli or in a stochastic way with age.

It's fundamental to consider the DNA in the context of chromatin, a more complex object that is not fully characterized just by the nucleotide sequence, there are other chemical modifications that occurs and can be taken into account.

A central role in the characterization of the epigenome is played by chromatin modulations. Histone composition and histone modifications work together with the binding of a large variety of other non-histone proteins to control both open (euchromatin) and closed (heterochromatin) chromatin states. The protruding tails of these histones can be modified by a variety of modifications such as methylation, acetylation, phosphorylation, and ubiquitylation, which take part in the determination of the transcriptional potential for a specific gene or a genomic region. It has been shown that DNA methylation is highly related to certain chromatin modifications. Regional chromatin structure seems to be linked with local DNA methylation thanks to the direct interaction with enzymes that modify DNA and histones. In the next section an introduction to DNA methylation will be given.

1.5 DNA Methylation

DNA methylation is the biological process by which methyl groups are added to the DNA molecule, as a modification of some of the bases. This modification is found almost exclusively on cytosines in the context of CpG sites, i.e. on cytosines followed by a guanine (5' → 3'orientation). The methyl donor is a molecule named S-adenosyl methionine (SAM).

The discovery of DNA methhylation is as old as the identification of the DNA as the genetic material. The modified cytosine was first discovered by Rollin Hotchkiss. Many researchers proposed the idea that DNA methylation could have a role in regulating gene expression, but the demonstration that DNA methylation was involved in gene regulation and cell differentiation came later in the 1980s . Now DNA methylation is well recognized as a major epigenetic factor influencing gene activities besides (and "cooperating" with) other regulators.

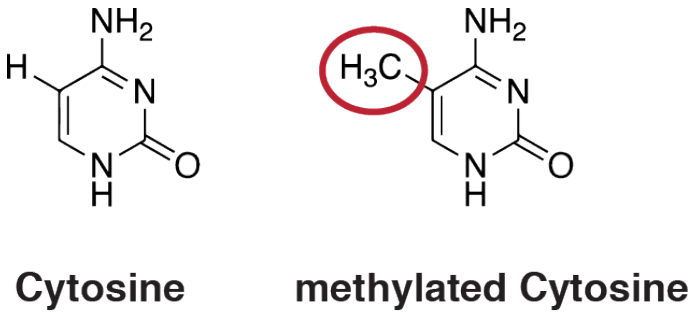

Fig. 1.4 ([23]) The methyl group is added on the 5' position of the pyrimidine ring of cytosines, that are then called 5-Methylcytosine (5mC)

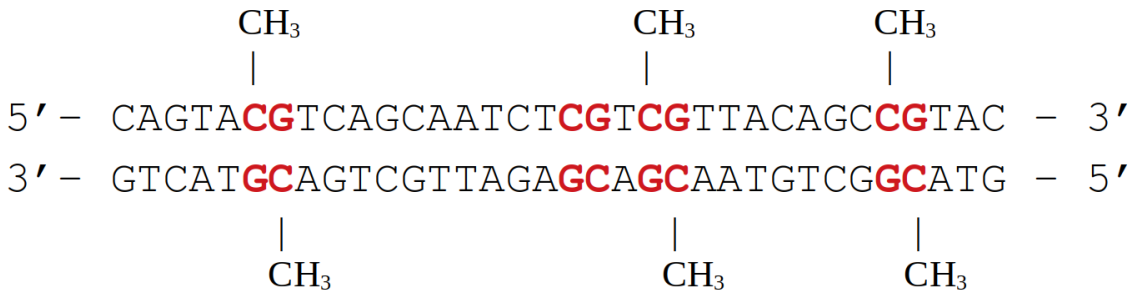

Fig. 1.5 Schematic representation of DNA methylation

Other types of cytosines methylation as in the context of CpA or CpNpG sequences (citosyne - any base - guanine) can be found in mouse embryonic stem cells, neurons, and plants, but are generally rare in mammalian tissues and less frequent than methylation in

the CpG context. Similarly adenine methylation has been observed in plant, bacterial, and also in mammalian stem cells DNA but has received considerably less attention.

In the human genome (haploid) there are about 29 million CpG sites. The proportion of methylated CpG dinucleotides varies slightly in different tissue types and is often around 60%-80%, so methylated cytosines represents about 1% of the total amount of bases.

The sequence symmetry of CpG dinucleotides permits to maintain the DNA methylation patterns through cell division, having led to the hypothesis that DNA methylation is part of the cellular identity and memory like the genome itself.

CpGs are underrepresented in the genome. The reason is that their mutation potential is about 10–50 times higher than other transitional mutations. As a result of this CpG dinucleotides occur only about one-fifth of the roughly 4% frequency that would be expected by simply multiplying the typical fraction of Cs and Gs (0.21*0.21). Here "expected" CpG frequency means the expected frequency supposing uncorrelated bases and certain frequencies for C and G, rather than the empirical observed one in the genome (≈ 0.01).

Despite this general trend, relatively CpG-rich clusters are present in the genome. They are called CpG islands, and are found in the promoter region and first exons (expressed part of a gene) of 65% of all genes. Promoters are regions about 100–1000 base pairs long placed before the transcription start site of a gene, to which proteins that start the transcription of that gene bind.

1.6 CpG Islands

There is not a formal definition of CpG island, usually they are identified as regions having an observed versus expected ratio for the occurrence of CpGs of at least ≈ 0.65 and a minimum size of 200-500 base pairs. Although this definition includes regions that have less CpGs than what would be expected in a random sequence of bases, this is acceptable because the actual observed versus expected ratio in most part of the genome is about 0.2. Depending on the definition, there are around 28.000-30.000 CpG islands in the human genome.

CpG sites in CpG islands are mostly unmethylated. This holds in all tissues and throughout all developmental stages. Moreover CpG islands often corresponds to an open chromatin structure and a potentially active state of transcription. The low level of methylation of CpG islands can explain the relatively high density of CpGs: they are less susceptible to deamination since CpG islands are mainly unmethylated from the germline. Other mechanisms that protect CpG islands from methylation are for example binding of transcription factors (proteins that controls the rate of transcription, by binding to a specific

DNA sequence) or exclusion of nucleosomes (DNA regions in which a nucleosome can't bind since there are DNA sequence patterns that are too rigid to form loops).

CpG islands have a role in promoting gene expression by regulating chromatin structure and the binding of transcription factors, and for this reason they have been evolutionarily conserved. In general DNA is wrapped around histone proteins forming small, packaged sections called nucleosomes. The more DNA is bounded to histone proteins (heterochromatin), the less permissive it is for gene expression. Instead CpG islands are less associated with nucleosomes than the other regions of DNA.

The richness of G and C in many transcription factor binding sites suggests that CpG islands are likely to enhance binding to transcriptional start sites, although they are often devoid of common promoter elements. So CpG islands improve the accessibility of DNA and ease transcription factor binding even if they often lack in common promoter elements.

When a CpG island in a promoter region is methylated then gene expression is silenced in a stable way. This ability of regulating gene expression through CpG island methylation is important for the realization of imprinting . Imprinting is the phenomenon by which a gene is expressed from only one of the two inherited parental chromosomes and their expression is determined by the parent of inheritance. Then, beyond imprinted genes, it was observed that DNA methylation of CpG islands regulates gene expression during development and differentiation.

It would be expected that CpG islands display tissue-specific patterns of DNA methylation since they can silence and enhance gene expression. Actually this is rare in CpG islands associated with transcription start sites, even though it can happen for CpG islands in gene body and intragenic regions. Instead, highly conserved patterns of tissue-specific methylation can be found in regions called CpG island shores, the 2 kilobase neighborhood of CpG islands.

As in CpG islands, CpG shores methylation has a high negative correlation with gene expression.

Anyway the role of CpG islands in regulating gene expression is still not totally clear. Methylation of CpG islands can compromise transcription factor binding, recruit repressive methyl-binding proteins, and stably silence gene expression. However they are rarely methylated, especially those associated with gene promoters. Further studies are needed to determine how much DNA methylation of CpG islands actually regulates gene expression.

1.7 DNA Methylation Mechanisms

It's possible to conceptually break the processes that modify the DNA methylation state in two categories: methylation writing, when a methyl group is added onto an unmethylated

cytosine, and methylation erasing, when the methyl group is removed. Enzymes that catalyze the transfer of methyl groups to DNA are called DNA methyltransferases (Dnmts).

1.7.1 Writing DNA Methylation

There are mainly three enzymes that directly catalyzes the addition of methyl groups on DNA: Dnmt1, Dnmt3a, and Dnmt3b. These enzymes share a similar structure but they have different functions and expression patterns.

Dnmt1 is probably the best known Dnmt, it's highly expressed in mammalian tissues including the brain and it's studied especially in the context of the nervous system. Dnmt1, unlike the other Dnmts, preferentially methylates hemimethylated DNA (locations where the methyl group is present in just one of the two strands). During DNA replication, Dnmt1 is found in the replication fork where newly synthesized hemimethylated DNA is formed. So Dnmt1 binds to the newly synthesized DNA and methylates it in the hemimethylated loci obtaining in that way a copy of the original methylation pattern present before replication. In that sense the activity of Dnmt1 is complementary to the other DNA replication enzymes, it allows to create a copy not only in the bases sequence but also in the methylation. The activity of Dnmt1 is not limited to the replication phase, it adds a methyl group whenever it finds hemimethylated DNA, so it can repair DNA Methylation. Therefore Dnmt1 is called the maintenance Dnmt, for its ability to maintain the original pattern of DNA methylation in a cell lineage. In fact it has been observed that the knockout (absence) of this enzyme in mouse results in embryonic lethality. Instead mouse embryonic stem cells lacking Dnmt1 remain alive but in vitro differentiation results in massive cell death. These findings demonstrate that Dnmt1 plays a vital role in cellular differentiation as well as in dividing cells.

Dnmt3a and Dnmt3b are extremely similar in structure and function. What mainly distinguishes them is their gene expression pattern: Dnmt3a is expressed almost everywhere, while Dnmt3b is poorly expressed by the majority of differentiated tissues with few exceptions. Dnmt3a and Dnmt3b are referred to as "de novo" Dnmt since they can add methylation into "naked" DNA, in the sense that where they are overexpressed, they don't show preference for hemimethylated DNA.

Also the knockout of Dnmt3b in mice is embryonic lethal, while Dnmt3a knockout resulted in death after 4 weeks. This suggests that Dnmt3b is required during early development, whereas Dnmt3a is required for normal cellular differentiation.

Finally Dnmt3L is a protein that lacks the catalytic domain (the region of an enzyme that interacts with its substrate to cause the enzymatic reaction [18]) present in other Dnmt enzymes. It's mainly expressed in early development and is present only in germ cells and thymus in adulthood. Even though Dnmt3L has no catalytic function by itself, it associates with the Dnmt3a and Dnmt3b to stimulates their methyltransferase activity. It has been

observed in mice that, consistently with its presence in early development and in germ cells, Dnmt3L is necessary for establishing genomic imprinting and for other functions (retrotransposons methylation, compaction of the X chromosome).

How the de novo Dnmts target specific DNA regions is still unclear, but several mechanisms have been proposed. A theory is that transcription factors can regulate de novo DNA methylation by binding to specific DNA sequence to either recruit Dnmts for methylation or protect from DNA methylation.

For example, CpG islands appear to be protected from methylation by transcription factor binding , in fact it seems that they are unable to maintain their unmethylated state when transcription factor binding sites are mutated, .

Several studies describe substantially two mechanisms that probably function together to set de novo DNA methylation. Dnmt3a and Dnmt3b can be recruited by specific transcription factors or simply they methylate all CpG sites across the genome that are not protected by a bound transcription factor.

1.7.2 Erasing DNA Methylation

Demethylation process can be either active or passive. Passive DNA demethylation occurs in dividing cells during DNA replication, when the inhibition or lack of Dnmt1 leaves the newly incorporated cytosine unmethylated.

Active DNA demethylation instead can occur in both dividing and non-dividing cells, but the process is more complicated. It involves multiple enzymatic reactions and enzymes to process the methylated cytosine in order to revert it back to a normal cytosine. In fact according to Moore et al. [17] much of the current scientific debate is probably motivated by this complexity.

1.8 Transcription

Transcription has always to be considered in the context of chromatin since the latter deeply influences the accessibility of the DNA to transcription factors and the DNA polymerase complexes. Chromatin remodeling, histone modifications and DNA methylation are closely linked epigenetic modifications that contribute to control gene expression through chromatin structure. As a proof it has been observed that mutations or loss of chromatin remodeling proteins lead to genome-wide modifications of DNA methylation patterns and altered gene expression programs.

DNA methylation may affect transcription in several ways. First, methylated DNA may inhibit the binding of transcriptional activators by sterical hindrance (non-bonding interactions), while other transcription factors are attracted by methylated target recognition sequences. Second, methylated DNA is bound by proteins that recruit transcriptional

repressors and chromatin remodeling complexes which action results in a repressive chromatin configuration.

In many cases DNA methylation occurs after changes in the chromatin structure and works as a way to lock this changes and keep the gene in a permanent inactive state. The fact that a CpG island or a gene regulatory element is unmethylated does not imply the transcriptional activity of the gene, but that the gene can be potentially activated.

Also the simple presence of methylation does not necessarily imply silencing of close genes. This happens only when a specific region of the promoter (that often spans the transcription start site) is hypermethylated.

1.9 Development and Differentiation

Methylation is fundamental for correct mammalian embryogenesis during which methylation levels change dynamically. During development and differentiation several cell-type-specific differentially marked epigenomes are created. DNA methylation patterns contribute to the definition of a cell identity. Thus, the human body contains several hundred different epigenomes still maintaining the same genome. As said before embryonic stem cells can survive in the absence of DNMTs, but their differentiation necessarily requires the presence of DNMTs.

When de-differentiation is forced with the aim of generating induced pluripotent stem cells, cells might conserve a slight preference to differentiate again into their cell-type of origin, suggesting that tissue-specific DNA methylation marks, that are not completely reset during the process, function as a memory for the cell.

In mammalian development there are two waves of genome-wide epigenetic reprogramming, in the zygote and in the primordial germ cells. Before implantation the genome in the zygote becomes demethylates, probably to initiate cellular differentiation. Most of the paternal genome actively and rapidly demethylated leading to the erasure of most paternal methylation marks, while the maternal genome remains methylated, although recent results suggests some active demethylation also in the maternal genome.

During implantation, where cells start to differentiate into different developmental lineages, DNA methylation levels are restored by de novo methylation. In this phase disruption of any of the DNA methyltransferases results in embryonic lethality.

Early stages of development are characterized by high epigenetic plasticity or reprogramming, thus environmental stimuli in this lapse of time can play an important role since they can lead to permanent changes in the patterns of epigenetic modifications. In vitro fertilization and the associated cell culture are associated with changes of DNA methylation patterns. In fact there are studies that show that imprinting disorders are slightly more probable in the context of assisted reproductive technologies.

1.10 Aging

There are different ways in which epigenetic changes and methylation are related to the aging process. First, the overall content of DNA methylation in the mammalian and human genome decreases with age especially at repetitive elements (patterns of DNA that occur in multiple copies along the genome). At the same time, distinct genes acquire methylation at specific sites including their promoters, a process that resembles the DNA methylation changes found in cancer. Aging is then associated to the phenomenon of epigenetic drift, the divergence of cells epigenetic characteristics as a function of age caused by stochastic changes in methylation.

It has been shown that is possible to predict the age of an individual with relatively high precision from DNA methylation, defining in this way the concept of epigenetic age. Those predictors are called "epigenetic clocks". The first robust multi-tissue epigenetic clock was proposed in Horvath [12], the model utilizes the methylation levels of a specific set of 353 CpGs across a wide spectrum of tissues and cell types and achieve a median error of ≈ 3.6 years.

Several diseases such as HIV infection and Parkinson's disease are associated with an increment of epigenetic age, while it has been observed that people with exceptional longevity show a decreased epigenetic age compared to their chronological age.

1.11 Other Methylation Functions

1.11.1 Gene Body Methylation

Gene bodies are commonly highly methylated, and, contrary to CpG islands, DNA methylation has been associated with enhanced gene expression. The reason could be the prevention of initiation of spurious transcription events, or maybe because it makes a phase of transcription, the transcriptional elongation, easier. The gene body is considered the region of the gene past the first exon, since methylation of the first exon leads to gene silencing, because the first exon often overlaps with a CpG island. However this is not observed in slowly dividing and non-dividing cells such as the brain.

In general how DNA methylation of the gene body contributes to gene regulation remains unclear.

1.11.2 Imprinting

In mammals, the maternal and paternal genomes are functionally different and both are required for normal development. There is a subset of genes that is asymmetrically

expressed from only the maternal or the paternal allele (the two corresponding variants of a gene) in all somatic cells of the offspring, these are called imprinted genes.

Generally these imprinted genes are displaced in clusters and the alleles show different marks in DNA methylation, histone modifications and other epigenetic features. Around 150 imprinted genes are known in mouse and man, and some more imprinted genes have been computationally predicted. Imprints are established in the gametes by Dnmt3a and Dnmt3l in a specific manner that depends on the parent of origin.

1.11.3 X Inactivation

In female mammals cells one of the two X chromosomes (at random) is inactive in order to achieve dosage compensation, this represents an excellent example of stable and heritable epigenetic modification that regulates gene expression. The inactive X chromosome is silenced thanks to its packaging into a transcriptionally inactive structure, the heterochromatin. DNA methylation patterns are established on the inactive X chromosome and contributes to maintaining it in such configuration.

1.11.4 Genome Stability

DNA methylation is fundamental in the maintenance of genome integrity by silencing of repetitive DNA sequences and endogenous transposons, a type of genetic component of DNA that are able to copy and inserting themselves into different genomic locations. In fact, it has been shown that the absence of DNA methylation leads to the reactivation of retroviruses during embryonic development.

Retrotransposons activity could be disabled also thanks to the increased mutation rate caused by methylation, leading in this way to a faster divergence of identical sequences.

Chapter 2

Multiscale Relevance

In this chapter we introduce the *multiscale relevance*, a measure for characterizing the variability of real values in a sample across different scales and consequently, for identifying samples that contain information about the generative process. This non-parametric and model-free measure was first introduced in Cubero et al. [7] as a tool for selecting "relevant" neurons by their observed activity, without requiring knowledge of external correlates. The motivations that led to the definition of *multiscale relevance* are rooted in information theory, in particular in a number of recent publications on criticality of efficient representations (Marsili et al. [16], Haimovici and Marsili [11], Cubero et al. [6]). The idea is to explore the application of this novel method to methylation data, with the aim of gaining insights in its functional role through an information theoretic characterization.

2.1 Definition

We introduce the method outlining the main ideas and the main steps. We refer to Cubero et al. [7] for a more detailed discussion of the method.

Given a sequence $\hat{x}$ of events that occur at times $t_1 \leq t_2 \leq ... \leq t_M$[1], we can define a discretization of these values into T bins B_s each one of size Δt: $B_s = [(s-1)\Delta t, s\Delta t]$ $(s = 1, 2, ..., T)$. We denote as k_s the number of values in B_s.

We think of the events being generated by an unknown process, that we call the generative process henceforth. Choosing a Δt allows to study the generative process at a fixed scale, but instead of Δt we could define an information theoretic measure of *resolution*:

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

[1] In Cubero et al. [7] these events were the spiking of neurons, now we can think of these events as the methylation or demethylation of a CpG.

This has the form of a Shannon entropy where we consider k_s/M as a probability (the probability that a value is found in bin B_s using Mats as unit of information. The "empirical" entropy $H[s]$ corresponds to the amount of information that one gains on a randomly chosen value t_i by knowing the bin B_s it belongs to. The idea is that $H[s]$ provides an intrinsic measure of resolution, contrary to Δt which depends on the specific dataset $\hat{x}$ of values. For example, for each set of values $\hat{x}$ there exist Δt_- and Δt_+ such that for all $\Delta t \leq \Delta t_-$ all bins contain at maximum one value, and for all $\Delta t \geq \Delta t_+$ all values are in the same bin. In the first case $k_s = 0, 1$ for all s and then $H[s] = 1$, in the second case $H[s] = 0$. These scales Δt_- and Δt_+ depends on the specific data x, whereas $H[s]$ provides an universal scale for resolution, that ranges between 0 and 1.

Given a certain binning corresponding to a Δt and characterized by a certain resolution $H[s]$, we can define a measure of dynamic/spatial richness of the process. Here the idea is that frequencies k_s are the only quantitative measure that can distinguish the process in two different bins B_s. Therefore, the richness of the dynamic/spatial behavior can be defined as the variability of different observed states, i.e. by the variability of the frequencies k_s. Again we measure this as an entropy and we call it *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.

$H[k]$ corresponds to the amount of information one gains on the index i of a randomly chosen value t_i by knowing the k of the bin it belongs to. We remark that $H[k] = 0$ in both the extreme situations discussed above: $\Delta t \leq \Delta t_-$ and $\Delta t \geq \Delta t_+$, then it reaches the maximum for an intermediate $\Delta t \in (\Delta t_-, \Delta t_+)$.

It's possible to calculate $H[s]$ and $H[k]$ for different values of Δt. 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 *multiscale relevance* (MSR). This allows to measure the dynamical/spatial richness at different scales, when the relevant scale may not be known.

By the data processing inequality we have that $H[k] \leq H[s]$ (since k_s is a function of s). However in the low resolution region generally $H[s] = H[k]$, since even if $p(s_i) = p(s_j)\ \forall i, j$, it is improbable to sample two or more outcomes s the same number of time if the sample size M is much larger than the number of possible outcomes.

2.2 Motivation

Formally, the multiscale relevance, or MSR, is a function that associates to a set of M real numbers a real number: $\mathbb{R}^M \to \mathbb{R}$. So it can be thought as a summary statistic, since the definition is independent from M and from the order of data.

Intuitively, it gives a measure of variability at different scales of a set of real values, but there are deeper theoretical motivations based on information theory.

Given a dataset $\hat{x}$ we could assign a label s to each of its values according to a labelling function $s()$ and then obtain a sample $\hat{s} = s(\hat{x})$. This is what is typically done in dimensionality reduction schemes and clustering algorithms to cope with high dimensional data in the deep undersampling regime, making some inference possible. As we have seen a sample $\hat{s}$ can be characterized by a *resolution* $H[s]$, that measures the level of detail of the description, and a *relevance* $H[k]$. In the absence of prior information, the only characteristic that could distinguish two different outcomes is their frequency in the sample. If two outcomes s_1 and s_2 occurs the same number of times $k_{s_1} = k_{s_2} = k$, then the distinction between outcomes s_1 and s_2 is based just on some pre-defined classification criteria. Then the larger the entropy of the frequencies $H[k]$, the larger the number of distinguishable outcomes, and then the more we can learn from data. In fact in both the two extreme situations where the description $\hat{s}$ of $\hat{x}$ is so coarse that all samples s_i are equal, or when it's so detailed that all samples are different, $H[k] = 0$ and actually we don't learn anything.

Starting from this intuition Cubero et al. [6] show that $H[K]$ provides an upper bound to the number of informative bits that the data contains on the generative process. Also $H[K]$ correlates with the number of parameters a model would require to describe properly the dataset, without overfitting (Haimovici and Marsili [11]). According to Cubero et al. [6] the frequencies $\hat{k}$ have the same role of minimally sufficient statistics in parametric model.

Marsili et al. [16] show that, for a given value of M and of the resolution $H[s]$, data that are maximally informative on the generative process are those for which $H[K]$ takes a maximal value. In the high resolution region (large $H[s]$, or small Δt when referring to the binning), the frequency distributions that achieve maximal values of $H[K]$ are broad. More precisely, the frequency distribution behaves as $m_k \sim k^{-\mu-1}$ where $-\mu$ is the slope of the $H[K]$-$H[s]$ curve. Indeed, μ quantifies the tradeoff between resolution ($H[s]$) and relevance ($H[K]$) in the sense that a reduction in $H[s]$ of one bit delivers an increase of μ bits in $H[K]$ (Cubero et al. [6]). In particular the celebrated Zipf's law ($m_k \sim k^{-2}$) [2] emerges at the

[2] In other terms, the frequency k_s of the most observed outcome s is n times the frequency of the nth most observed outcome

optimal tradeoff between resolution and relevance, since $H[k]+H[s]$ reaches the maximum when $\mu = 1$.

Thus Marsili et al. [16] suggest to probe the system on "critical" variables, since they provide more information on the system's behavior.

These arguments have been shown to be useful for example in characterizing the efficiency of representations in deep neuronal networks (Song et al. [19]) and for identifying relevant sites in proteins (Grigolon et al. [10]).

Hence MSR was designed to summarize how much information the data by itself contains on the generative process at different resolution levels. This is what was argued for the *relevance* $H[k]$, but in this case it's measured for different scales, since we ignore a priori which one is the more relevant for the observed system.

Therefore given several datasets $\hat{x}_i$, we could use MSR to detect the ones that are expected to have information on the hidden process, or from a different point of view, the data that is probably linked to a non trivial hidden behavior that does not only consists in noise.

More simply it can be viewed as a new measure of variability, characterizing samples with broad and non-trivial distribution across a broad range of scales.

Advantages of MSR are that it's non-parametric and fully featureless since it uses only the values of a sample x without resorting to any a priori covariate. This is remarkable because at the end the objective of an analysis could be to find a link between the process behind x and other known covariates.

2.3 An Application Example

The article *"Multiscale relevance and informative encoding in neuronal spike trains"* (Cubero et al. [7]) shows a successful application of MSR to neural activity data. Neural data consists in the recordings of time stamps at which a neuron "fires", for several observed neurons. In particular the experimental data was about a neural ensemble of about 800 neurons of a freely-behaving rat exploring a square area.

When applied to this data it was found that neurons having low MSR tend to have low mutual information with the correlates that are believed to be encoded by the region of the brain where the recordings were made. The opposite was not true, i.e. there were neurons characterized by an high MSR but unrelated to the rodent's "spatial" behavior, the reason could be that those neurons were related to other unobserved covariates.

So MSR was proposed as a measure to rank and select neurons for their information content without the need of any a priori covariate.

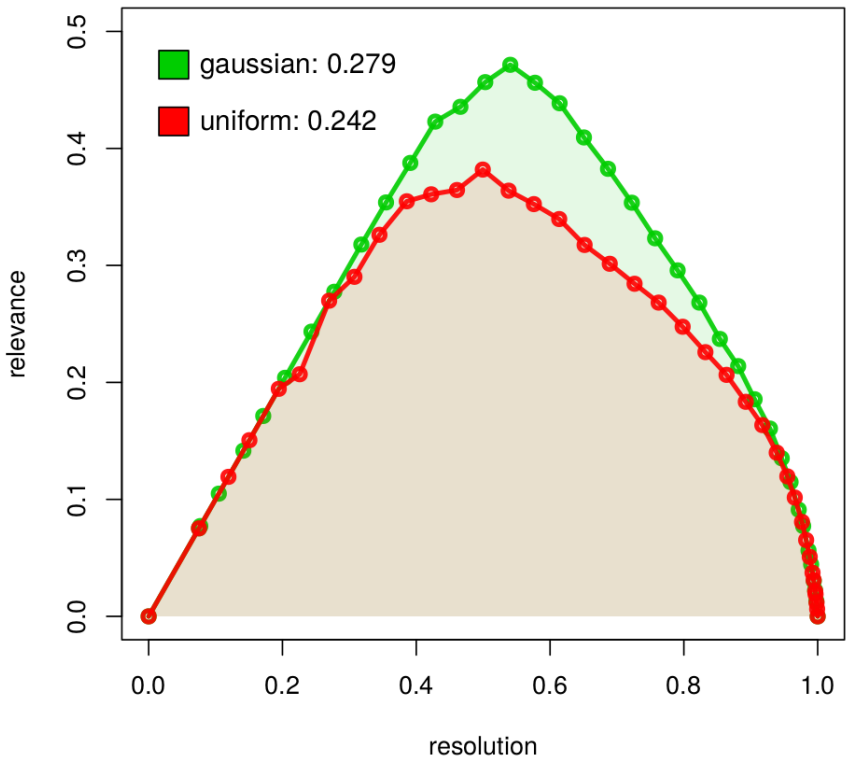

Fig. 2.1 *resolution-relevance plot for two samples of* 10000 *points generated respectively from a uniform distribution and a Gaussian distribution*. Each point in the plot corresponds to a different Δt. As expected the MSR for the second sample is higher, since compared to a uniform sample the density of points in different regions is more variable.

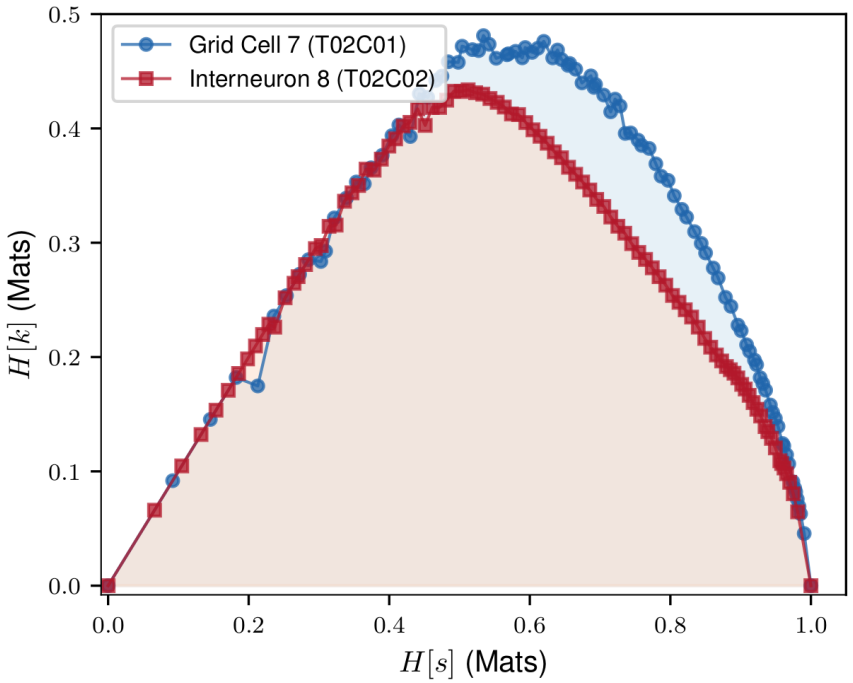

Fig. 2.2 (Cubero et al. [7]) *Proof of concept of the MSR as a relative information content measure*. The curves traced are relative to two different neurons, the one that traces the higher curve is also the one that is more related to the "spatial behavior" of the mice.

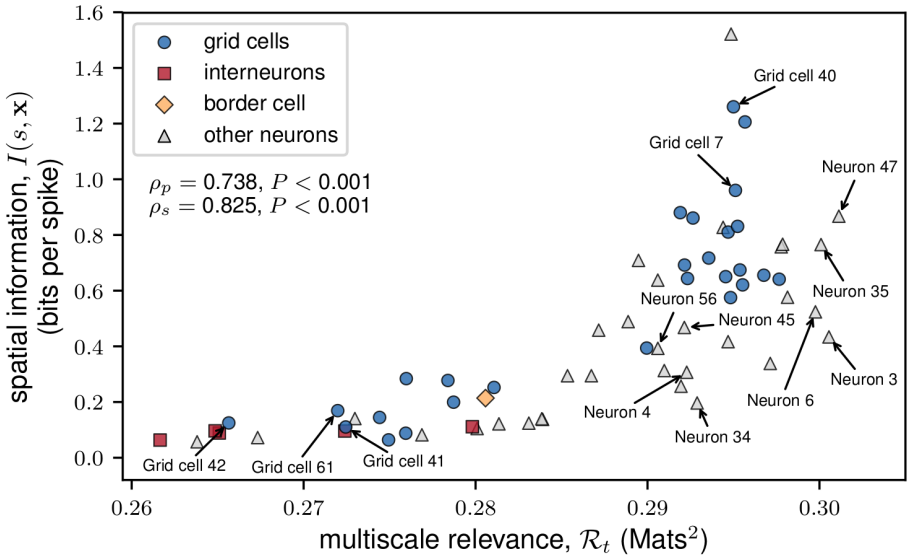

Fig. 2.3 (Cubero et al. [7]) This scatterplot shows MSR was able to separate neurons that contains no spatial information from the ones that could.

2.4 Computation Details

In the definitions some details about the actual computation were left unspecified.

In the binning when an interval $[a,b]$ is divided into T equally sized bins, the extremes a and b can be set a priori or they can be set to respectively the minimum and the maximum of the set of values (the smallest interval that contains all the values). It can be empirically verified that if M is large enough, then the choice of the interval is not determinant[3].

The number of points of the $H[s]$-$H[k]$ plot, that corresponds the number of different Δt for which the binning are built, influence the precision of the area computation. We will often use a value of at least 50 in this work, since a larger number seems not to significantly influence the result.

For each binning the bin size Δt (or equivalently the number of bins $T = (a-b)/\Delta t$) is not chosen arbitrarily. First a preliminary binning is made with a Δt_0 (corresponding to a T_0) small enough to let each bin contain at maximum one value[4], then each of the other are performed choosing as Δt a multiple of Δt_0. Multiples of Δt_0 are not always divisors of T_0, so the adopted solution was simply to choose only Δts for which the remainder was small and cut the length of the last bin. This preliminary binning is done in order to make the algorithm more efficient, and for another reason that we will see in the next chapter. If Δt_0 is small enough the approximation is valid.

We have observed that for very small values of M (approximately less than 50) MSR is unstable in the sense that it's too sensible to the position of each of the values and on the particular choice of bin sizes.

2.5 Statistical Analysis on random data

It's useful to study the statistical behavior of MSR, in order to have the possibility to check at least in an approximate way the significance of a given result.

First, there is a maximum value that MSR can assume that (slightly) depends on the sample size M. An upper bound of this maximum can be calculated solving the continuous counterpart of the optimization problem in which $H[k]$ has to be maximized at a fixed $H[s]$:

$$\begin{array}{ll} \max_{m_k} & H[k] \\ \text{subject to} & H[s] = c, \quad \sum_k k m_k = M \end{array}$$

It's also possible to obtain a lower bound of this maximum, forcing m_k to be Poisson variables with mean $\bar{m}_k$, and maximising the expected value of $H[k]$ with respect to $\bar{m}_k$, at fixed expected values of $H[s]$ and sample size M (see Haimovici and Marsili [11] for more

[3] as long as it includes all the values

[4] In this case the computation of $H[s]$ and $H[k]$ is not necessary, since the point ($H[s]=0$, $H[k]=0$) is known a priori together with the point ($H[s]=1$, $H[k]=0$)

details). To give an idea we can say that the maximum for MSR is ≈ 0.3 for a sample size of at least 100.

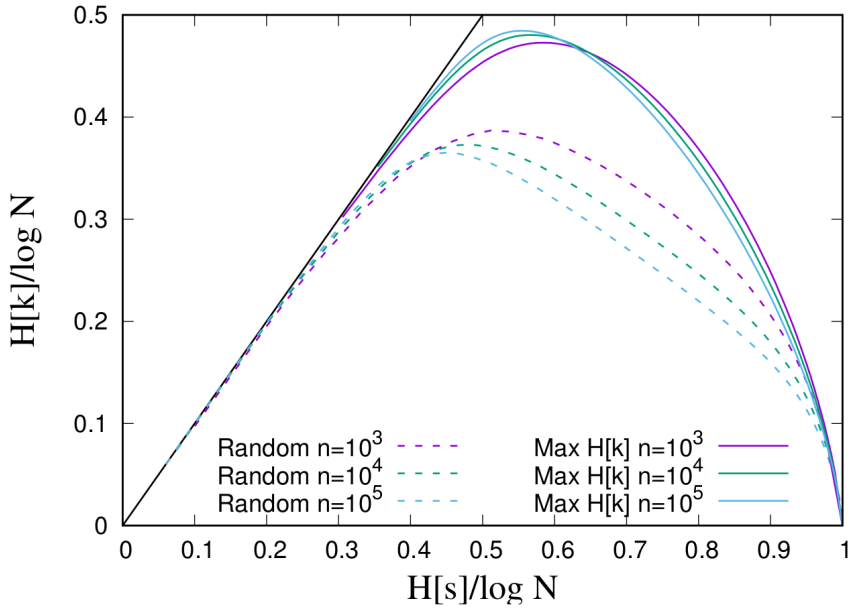

Fig. 2.4 (Cubero et al. [6]) $H[k]$-$H[s]$ curves for structure-less samples (dashed lines), for $M = 10^3$, 10^4 and 10^5. For structure-less samples, the averages of $H[s]$ and $H[k]$ were computed over 10^7 realisations of random distributions of M balls in L boxes, with L varying from 2 to 10^7. Each box corresponds to one state $s = 1, ..., L$ and k_s is the number of balls in box s. The full black line represents the limit $H[k] = H[s]$ which is obtained when $m_k \leq 1$ for all k. This limit is imposed by the data processing inequality. The full coloured lines instead are lower bounds for maximal MSR curves as computed in Haimovici and Marsili [11].

We would like to know the behavior of MSR for unstructured samples with a finite domain, i.e. for uniformly distributed samples. In figure 2.5 we can see the MSR for 10^4 samples simulated from a uniform distribution, for a varying sample size M. In the plot we used $\log_{10} M$ in order to better show the relation with MSR.

First of all we observe a negative correlation of MSR with the sample size, in particular it's almost linear with respect to the logarithm $\log_{10} M$. Then the variance decrease as $\log_{10} M$ increases. Points that corresponds to a small M are probably to be discarded since the MSR is too much "instable" in the sense that depends too much on computation details and on single values of the sample (MSR in those cases can assume also very small values), so we will focus on samples that have at least 100 elements, for which MSR has an acceptable noise.

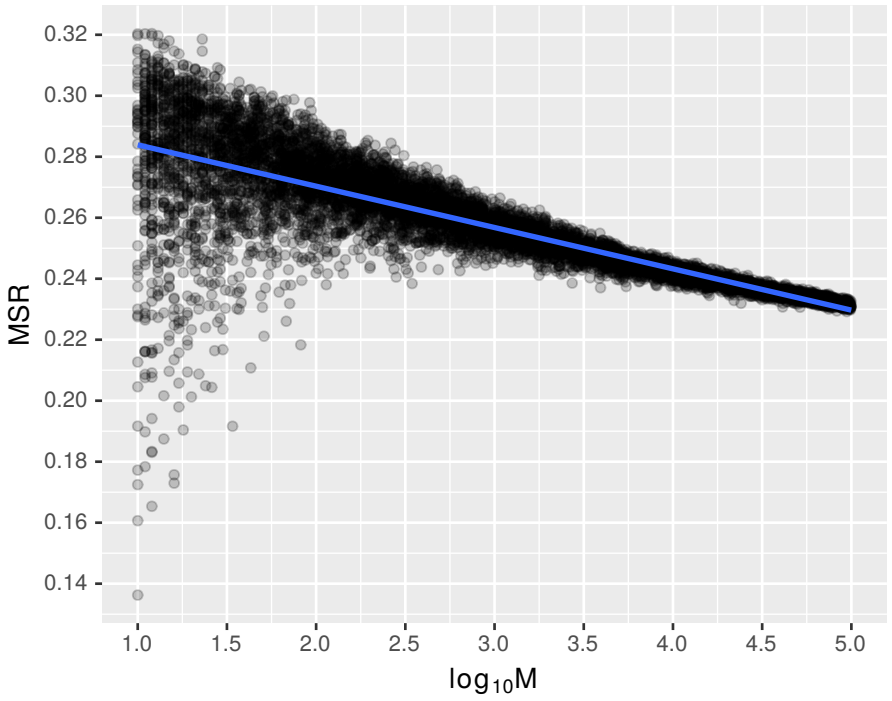

Fig. 2.5 Each of the 10^4 point represent the size M and the MSR of a random sample from a uniform distribution. In particular $log_{10}M$ was sampled from a uniform distribution in the interval $[1,5]$. The blue line is the result of a linear regression fit.

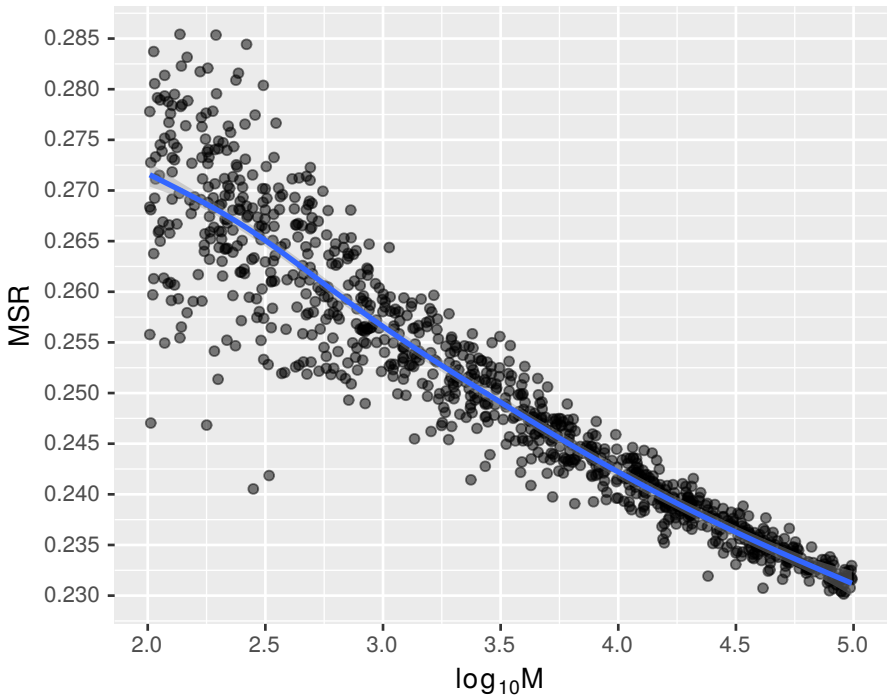

Fig. 2.6 Each point represent the size M and the MSR of a random sample from a uniform distribution. This time the number of points was reduced and the domain of M was limited to $[10^2, 10^5]$. The blue line is the result of a generalized additive model fit.

Chapter 3

Application of MSR to Methylation data

3.1 Methylation Data

As the genome is the complete information encoded in the DNA sequence of bases, the methylome of a particular cell of an organism contains the information about the methylation state of each cytosine of the genome. Thanks to next-generation sequencing techniques it's now possible in a certain measure to access this knowledge. In general sequencing means to determine the structure of certain long biomolecules like DNA, RNA, and proteins.

In particular Whole Genome Bisulfite Sequencing (WGBS) is a sequencing technique that permits to get methylation information genome-wide and at single nucleotides resolution. For this purpose WGBS is the most accurate technique but it's also expensive. An alternative is Reduced Representation Bisulfite Sequencing (RRBS) that is a cheaper technique but only covers CpG rich areas of the genome. In this work we will use only WGBS data since our aim is to apply the same analysis to any area of the genome without any bias towards CpG rich areas (that are probably the one that are believed to be "functional"). Furthermore we are only interested in methylation of cytosines of CpG sites, even if methylation is also possible for cytosine in other contexts or also for adenine.

The methylome is not the same for each cell, as discussed in the first chapter, human body has several differentially marked epigenomes, that differ also in the methylation patterns, depending on the specific cell function or tissue. Moreover there is heterogeneity also between cells of the same tissue caused by the stochasticity that characterize the DNA methylation processes and by epigenetic drift. So typical sequencing techniques allow us to extract information on a bulk of cells, getting in this way "averaged" data. In particular in WGBS data we have for each CpG site i a sample of n_i reads with the information on the proportion p_i of reads for which that CpG was found methylated. The more reads for each CpG the better, since it allows us to know the proportion of methylated CpGs with a

higher resolution. In fact the average number of reads for each site is taken as a measure of quality of sequencing data in general.

Formally, WGBS data consists of a vector $\bar{v}$ as long as the number of CpG sites in the DNA, in which each component v_i is a tuple (chromosome, position, strand, reads, proportion) containing information on a CpG site. The first three components indicate its genomic location, respectively the chromosome, the position[1] and to which strand the position refers (+ or -). The last two components indicate the number of reads and the proportion of methylated ones. For each location there are both the + and - strand variants since to each CpG in one strand corresponds one in the other. Elements (or rows) for which the number of reads is 0 are considered as missing data.

For our purposes WGBS data from ENCODE database is used (Consortium et al. [3]). ENCODE is a public research project which aims to identify functional elements in the human genome, its database contains several reference epigenomes for different tissues and types of cell (for both human and mice) (Consortium et al. [4]).

	chr	pos	strand	reads	prop		chr	pos	strand	reads	prop
1	chr1	10468	+	1	1.00	*58304907*	chrY	56887579	+	9	0.67
2	chr1	10469	-	0	0.00	*58304908*	chrY	56887580	-	13	0.54
3	chr1	10470	+	3	1.00	*58304909*	chrY	56887581	+	9	1.00
4	chr1	10471	-	0	0.00	*58304910*	chrY	56887582	-	13	1.00
5	chr1	10483	+	3	0.33	*58304911*	chrY	56887700	+	6	1.00
6	chr1	10484	-	0	0.00	*58304912*	chrY	56887701	-	7	0.71

Fig. 3.1 *Example of first and last rows of a table representing a WGBS of stomach tissue from ENCODE.* The data we are interested in is a table in which three columns identify the position of the cytosine (chromosome, position and strand) and the other two give information about the methylation state. This table contains the position of all and only CpG sites in the genome. The other two columns (reads and prop) indicate the total number of reads for a certain cytosine and the proportion of methylated ones. The presence of both methylated and unmethylated reads for the same cytosine is due to the heterogeneity in the sample of cells that are sequenced, and also to sequencing errors.

3.2 Methylation data representation

If we want to apply MSR to methylation data, we have to represent it as a set of values, in particular a set of positions indicating the methylated loci in the genome.

The idea is to represent the genome as a binary string ($\{0,1\}^n$), where the ones represent a methylated cytosine, or equivalently as a set of natural numbers indicating the position of all methylated cytosines.

[1] usually it refers to the position of the C of the CpG

Let's consider a sequence of nucleotides in the genome, the following are two ways in which it's possible to represent a methylome as a binary string. The first method consists in replacing each base with a 1 if it's methylated and with a 0 otherwise (Fig. 3.2). The string obtained in this way will be of the same length of the genomic region. Obviously each base that is not a cytosine of a CpG site will be 0.

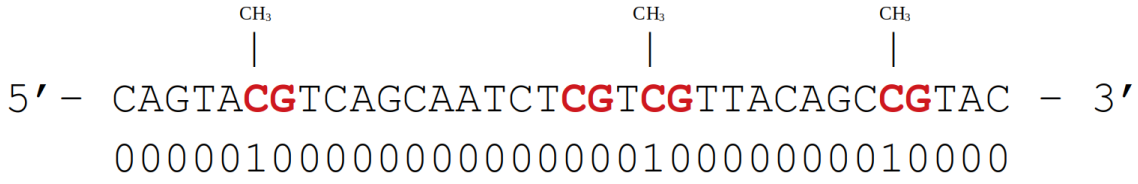

Fig. 3.2 A first way to encode methylation information into a binary string is to place a 1 where there is a methylated cytosine, a 0 otherwise This is an example on a hypothetical genomic region.

The second method is to consider only the cytosines in CpG sites, making them contiguous in the string (Fig. 3.3). Then in the same way a methylated cytosine is associated to a 1 and 0 otherwise. In this way the string can potentially be 1 in each position.

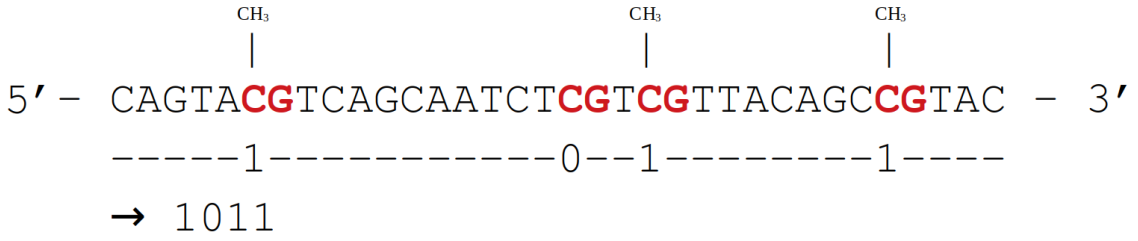

Fig. 3.3 A second way to encode methylation information into a binary string. This time only CpGs sites' cytosines are kept.

The second representation is more compact, since we throw away the redundant zeros, at the cost of losing information on the spatial distribution of CpG sites along the genome. We will often prefer the second way to represent the methylome.

We have also the possibility to study with MSR other genomic features besides methylation, like the positions of CpG dinucleotides in the genome, obviously this is not as interesting as methylation since it's not an epigenetic feature and it's (almost) the same for all cells (Fig. 3.4).

In order to represent the WGBS data in this way we need to apply some transformations first, since we don't actually know the methylome of a given cell, we can only access average information on each CpG.

The first transformation that is applied to the table is to merge the reads of the two strands, since we assume that the methylation of CpG sites is symmetrical, i.e. if a CpG

5′– CAGTA**CG**TCAGCAATCT**CG**T**CG**TTACAGC**CG**TAC – 3′
0000010000000000010010000000010000

Fig. 3.4 In this figure it is shown a way to encode the positions of all CpG sites' cytosines in a binary string.

site is methylated then it is also the corresponding GpC on the opposite strand (the opposite phenomenon is called "hemimethylation")[2] . It's also reasonable to consider methylation mainly as a phenomenon independent of the strand, as the enzyme methyltransferase Dnmt1 detect hemimethylated sites and add the methyl group to the other strand. So in this work we will ignore strand information.

The next step is to transform a proportion of methylated reads into a binary value (Fig. 3.5). A simple way is just to consider a site methylated if the proportion of methylated reads is bigger or equal 50%, else it's considered as unmethylated[3]. With this binary approximation, the information about the confidence of the methylation state of each site, or from a different point of view, the information about heterogeneity in cells methylation, is lost. This loss of information is mitigated by the fact that in most cells the proportions are in most cases near 1 or 0.

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

Fig. 3.5 *Example of methylation data processing (without chromosome information)* . In the second table strand information is ignored, in the third a binary value is used to represent methylation state.

There can be a percentage of CpG sites with no reads, those sites are considered as missing data. If the percentage of missing data is small enough (below 10%) in a certain binary string, it's still possible to calculate MSR at the cost of getting noise as we will see.

[2]This is supported by the fact that in the majority of CpG sites (those in which reads for both strands are available) the proportion of methylated reads is coherent in the two strands.

[3]In the specific case of a proportion of exactly 50% we can choose to consider that site as a missing data

3.3 MSR for methylation data

Having transformed the WGBS data in a binary string, we can now apply MSR. MSR can be applied to a set of real numbers, in a similar way we can apply it to the binary strings representing methylation by calculating the MSR on the set of discrete positions (indexes) of the ones in the string. Notice that we can also do it for the indexes of the zeros in the binary strings.

There are two fundamental differences between this data and a random sample of real numbers (the data to which MSR was design to be applied). First the values are constrained to be natural numbers since we are representing indexes in a sequence. This implies there is a minimum scale Δt_0 (that correspond to the minimum possible bin size) at which we can evaluate the generative process, that correspond to the maximum resolution scale by default. Then at each bin of size Δt_0 there can't be more than one value, this means that formally the data can't be regarded as the result of several independent samples from a discrete distribution. Anyway we will apply MSR irrespective of this, analyzing another time the behavior of this statistic applied to "random" data.

Given strings of length n we can consider as random data the result of n samples from a Bernoulli distribution with parameter p. So for a given n we can observe the behavior of MSR as we vary p, in a similar way of what we did with random samples from a uniform in the previous chapter.

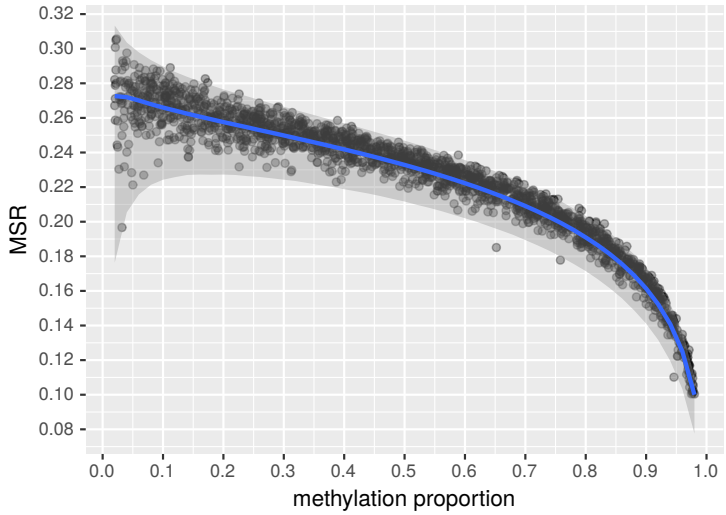

Fig. 3.6 For 2000 times a random binary string of length 1000 was sampled sampling each component from a Bernoulli with parameter p (p was sampled randomly from a uniform distribution in $[0,1]$). Then each point in the scatter plot represents the couple (MSR, proportion of ones) for a given random binary string. The blue line is the estimated median while the grey area represents the 99% confidence interval of the simulated data. These statistics were empirically estimated sampling 10^4 binary strings with a wide range of proportion of ones.

We can see from figure 3.6 that MSR rapidly decrease as the proportion of ones increases. This is not analogous to the negative correlation with $\log M$ observed for samples from a uniform distribution, instead is caused by the "saturation" that those strings with a high proportion of ones encounter. When there is a one in almost all positions, then many values will share the same k of its bin (the maximum), especially for small Δts, causing the relevance $H[k]$ to be small.

3.4 MSR related statistics

We can "adjust" the MSR taking into account the density of ones of a given strings aiming to a measure that gives an idea of the significance. For example we can subtract from the MSR the expected value (or the median) for the value of the density of that binary string, making it unbiased.

Let's define as $X(l,d)$ as the random variable representing binary strings generated in the way described above (l independent Bernoulli trials such that the proportion of successes is d). Then if x is a particular binary string characterized by a length l and a density (proportion of ones) d, then we define as $X|x := X(length(x), density(x))$ the random variable X conditioned on the characteristics of x. Now we can define $residual(x) := MSR(x) - E[MSR(X|x)]$. In this way $residual(x)$ represents how much more or less is MSR(x) than the expected. But this does not coincide with a measure of significance, moreover the variance is not the same for strings with a different density. We can use a cumulative distribution function to tell more about the significance of a result, in our case we give the following definition: $ecdf(x) := P(MSR(X|x) < MSR(x))$. $ecdf(x)$ tells how much probable is to find randomly an MSR value that is less than MSR(x).

Those statistics can be obtained with simulation, this is the reason why we call it empirical-cdf (ecdf) instead of cdf. As a consequence ecdf will be an approximation of the real cdf, for example there are strings x for which $ecdf(x)$ is exactly 0, since $MSR(x)$ is lower than every other simulated value. This also imply that those statistics are computationally hard to compute, since for a given binary string of length l and density d, we should simulate enough values of $MSR(X(l,d))$.

So for a given binary string x, representing the methylation state of a certain genomic area, we add to the MSR other two related statistics. Moreover we can apply those statistics also considering the positions of the 0s instead of the 1s when computing MSR. To distinguish them we will use the following nomenclature: MSR_1, $residual_1$, $ecdf_1$, MSR_0, $residual_0$, $ecdf_0$.

3.5 Dealing with missing data

It can happen that for certain CpGs there are no reads, or however their number is considered not enough to determine the methylation state. In this case these values are considered as missing, but still we want to calculate MSR. So the question is how we can approximate MSR in presence of missing data and how much the error is influenced by the number of missing elements.

A solution can be to replace each of these missing values with a good guess. A trivial but ineffective solution is to replace each of the missing values with a 1 or a 0, this leads to biased results. The solution we propose is the following: each time we are calculating the k of a bin, we replace the missing elements in the bin according to its proportion of 1s. What we do is to calculate the proportion p of ones, then the total count is computed as $\hat{k} = p\Delta t$ with Δt being the actual bin size. Since $\hat{k}$ may not be a natural number, it is rounded[4].

There can also be bins in which a large proportion of values is missing, in these cases it does not makes sense to replace the missing values. The alternative is to simply discard such bins whose missing data fraction is greater than a certain threshold.

We can then evaluate through simulation how much error we commit by resorting to this technique for dealing with missing data. We simulated several binary strings for different generative processes, calculating for each one the actual MSR and the MSR after removing a certain percentage of elements at random. For a missing data proportion of 10% on strings of length 1000, the root mean squared error was always less than 0.01, and often around 0.005. We noticed that the error was related to the variance of MSR of a certain generative process.

If we discard only the CpG sites with no reads, the missing data proportion is less than 5% for most of WGBS data we used. So we expect to have an acceptable noise when calculating MSR for most of genomic regions.

3.6 MSR distribution across genome

The idea is to apply the MSR to different genomic regions in order to study if MSR can capture those ones that show interesting and more rich methylation patterns, or the ones that have more information in the sense we discussed in the previous chapter. At the end we would like to find relationships between MSR and other features characterizing a certain genomic region.

We obtain regions subdividing the entire methylome represented in a WGBS of a certain tissue or cell type in fragments of a certain fixed number of CpGs. Those fragments contain the same number of CpG sites but the length in terms of nucleotides of the relative

[4] In order to obtain less biased results we choose to do a "stochastic rounding" of $\hat{k}$: $\hat{k} = \lfloor k \rfloor + \text{Bernoulli}(k - \lfloor k \rfloor)$.

region is variable. Keeping fixed the number of CpGs instead of the length in terms of nucleotides guarantees that there is enough data to calculate certain statistics as MSR, which also result easier to compare between the fragments.

As a first example we use stomach WGBS data from ENCODE, and fragments of length 1000 (CpGs). Considering that there are $\approx 2.9x10^7$ CpGs in human genome this results in a dataset of $\approx 29,000$ fragments. For each fragment we calculate basic features as the methylation rate (the mean of the proportion of methylated reads for each CpG) or the number of nucleotides that are embraced from the first to the last CpG of the fragment. Then we calculate MSR related features based on the transformation of the fragments in a binary string representing the CpG list: MSR_1, MSR_0, $ecdf_0$, $ecdf_1$, $residual_0$, $residual_1$.

From the total amount of fragments we discard the ones having more than 10% of missing data, having a methylation rate outside the interval (2%, 98%) or also embracing a very large genomic range ($> 3x10^5$).

First we observe from 3.7 in which we plot the methylation rate vs the MSR that the distribution of the points is actually different from the random one. In particular the 99,9% confidence interval for random binary strings highlights that a consistent fraction of points is outside this interval, meaning that there are several fragments for which a certain value of MSR for a certain proportion of 1s is almost impossible to obtain randomly, as also shown in 3.9. Notice that there are many points that are in the "saturation" region, where their MSR_1 result limited but still above the average. There are more points that have a remarkably low MSR_0 or MSR_1 than the opposite, in particular for an intermediate methylation rate.

From 3.10 instead we can see that MSR_1 and MSR_0 have a different distribution, the first seems a normal while the second is broader and bimodal. This is a first clue that these two quantities are not related at least in a trivial way.

It's interesting to analyze the correlations between all these features as it is shown in 3.11. As in the case of random data MSR_1 and MSR_0 are correlated to the methylation rate, but in this case it also happens for residuals and the ecdfs. Moreover all these significance measures are highly correlated between themselves, and we observed this is also true in WGBS data of different cell types. In a certain sense this tells us that when MSR_1 has a "significant" value then this also happens for MSR_0, though there is only a slight linear correlation between them.

We also calculated for each fragment the MSR of the 1000 genomic positions of CpG dinucleotides, and then we compared the distribution with random samples of 1000 uniform distributed points. Notice that in this case we are no more dealing with binary strings of a fixed length (1000), instead we are using sets of $M = 1000$ discrete positions, that correspond to much larger strings (see 3.4). Figure 3.12 shows that this time we get much higher values than expected, this is even more marked in fragments of 10,000 CpGs **??**. Although positions of CpG dinucleotides do not depend on methylation (they are not

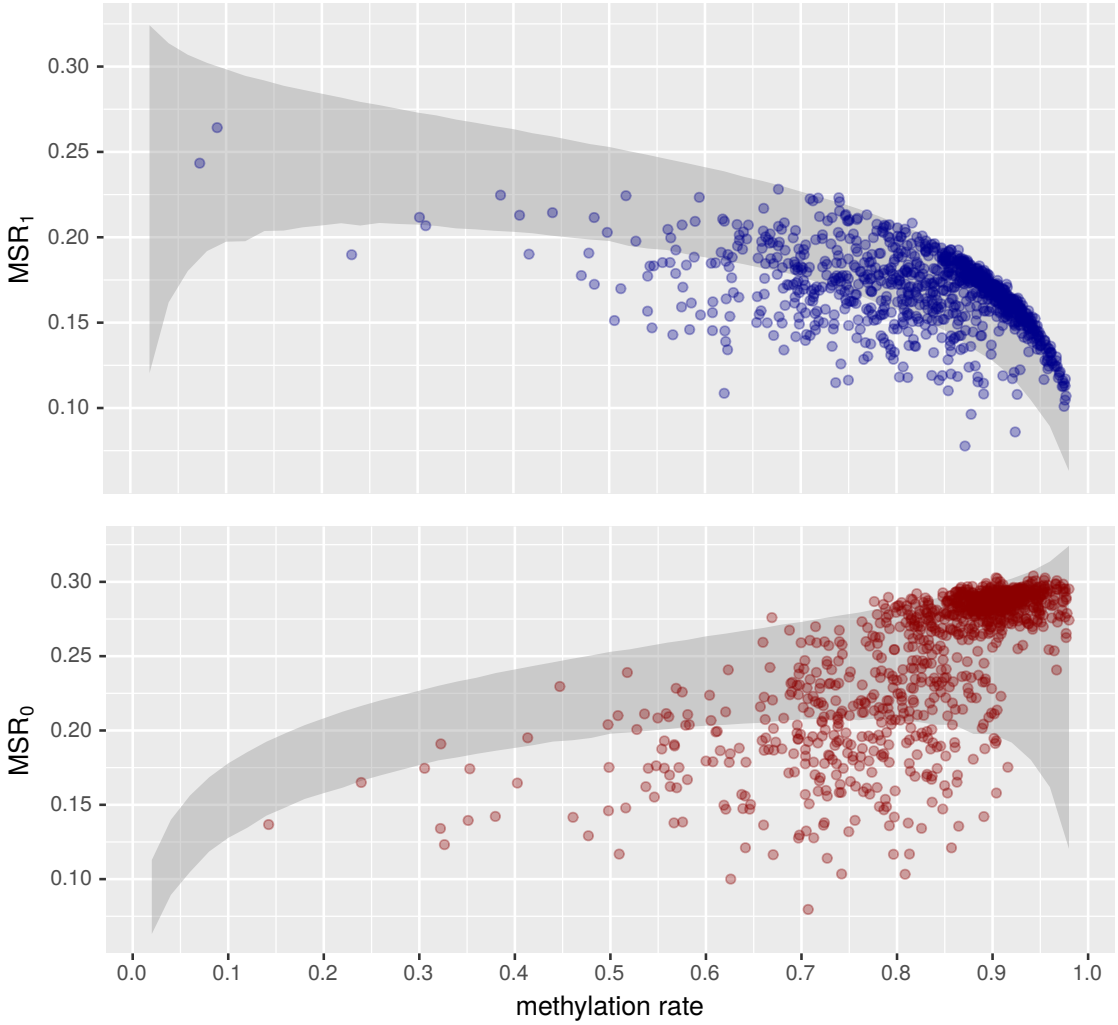

Fig. 3.7 In the two plots each point represent the methylation rate of a fragment with its MSR_1 and MSR_0. The gray area covers the 99.9% confidence interval for random binary strings (the same in 3.6), in the second plot it's mirrored with respect to the y axis since for MSR_0 we consider the proportion of 0s instead. In order to make it more understandable only 1000 randomly sampled points are shown for each of the two plots. In this plot we improperly called "methylation rate", the proportion of ones after the discretization, while it's usually the mean of the proportion of methylated reads. Figure 3.8 shows that the two quantities are not exactly the same.

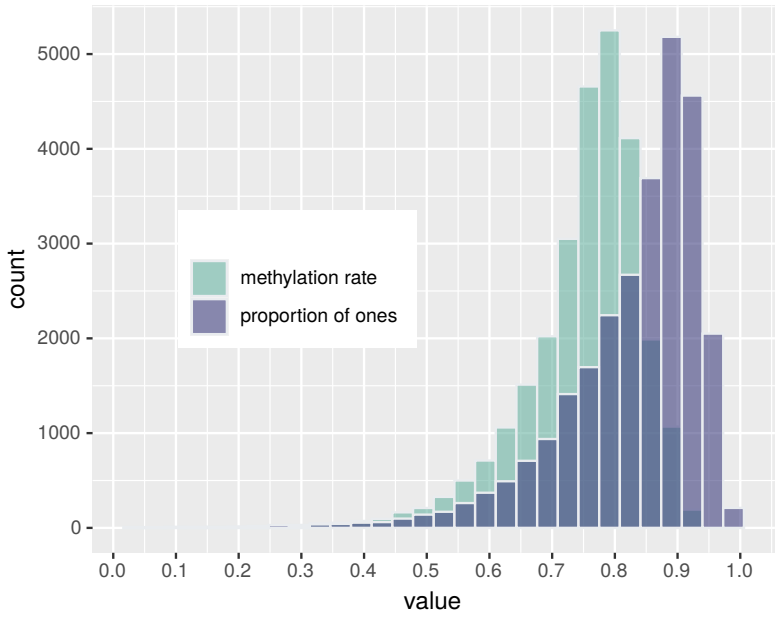

Fig. 3.8 The methylation rate for each fragment is the mean of the proportion of methylated reads for each CpGs, while the proportion of ones is the proportion of methylated positions after the discretization performed in order to calculate the MSR. Though the discretization results in a inflation of the rate, the two quantities are highly correlated ($r = 0.98$)

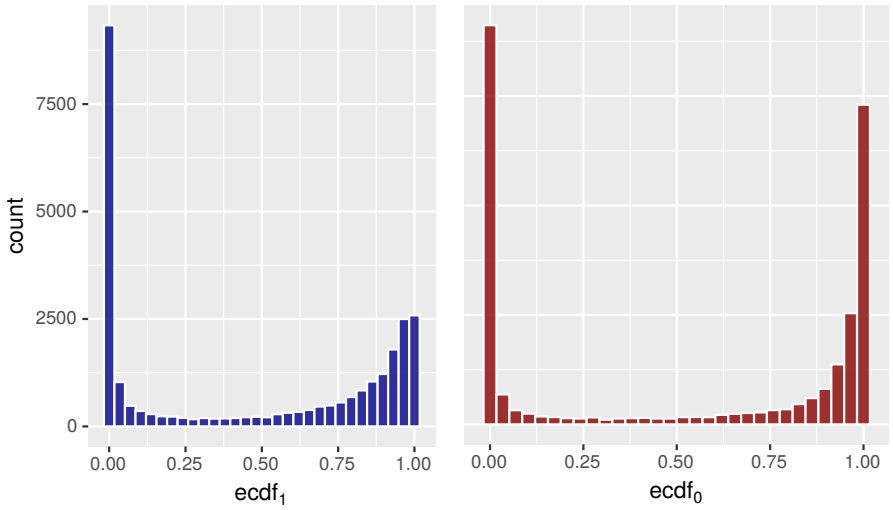

Fig. 3.9 *Histograms of* $ecdf_0$ *and* $ecdf_1$. These two histograms show the distribution of respectively $ecdf_1$ and $ecdf_0$ for stomach fragments of 1000 CpGs. In both cases peaks in 0 and 1 denote the presence of several fragments with non-trivial methylation patterns as they assume MSR_0 and MSR_1 values that are almost impossible to obtain randomly.

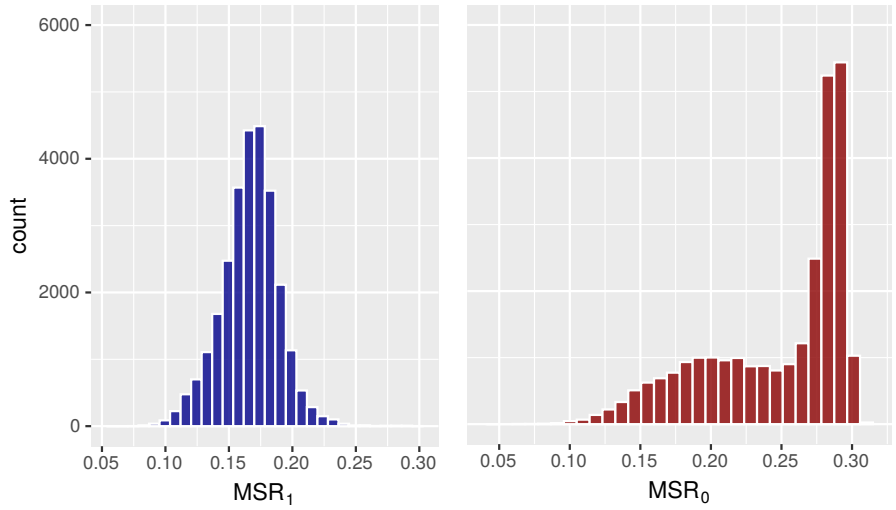

Fig. 3.10 *Histograms of MSR_0 and MSR_1.* These two histograms show the distribution of respectively MSR_1 and MSR_0 for stomach fragments of 1000 CpGs. The distribution of MSR_1 resembles a normal, while the distribution of MSR_0 has a more interesting shape with two modes.

an epigenetic feature), it is interesting to note that these sets of 1000 positions have almost never a trivial distribution in the genome.

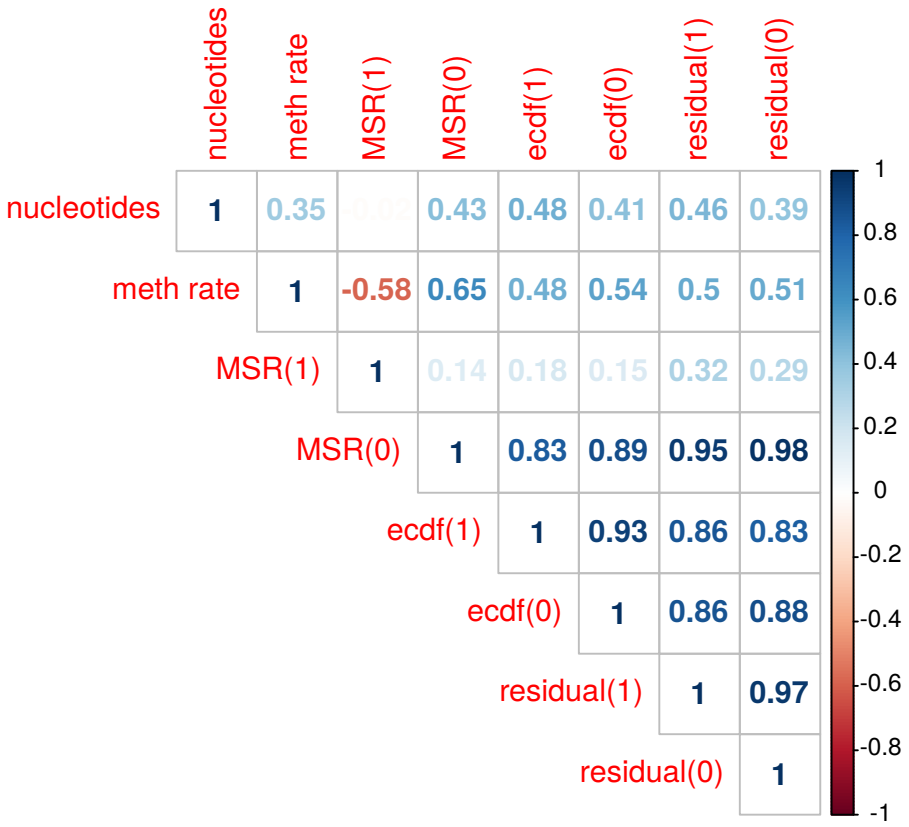

Fig. 3.11 *Pearson correlation coefficients between several features of stomach fragments*

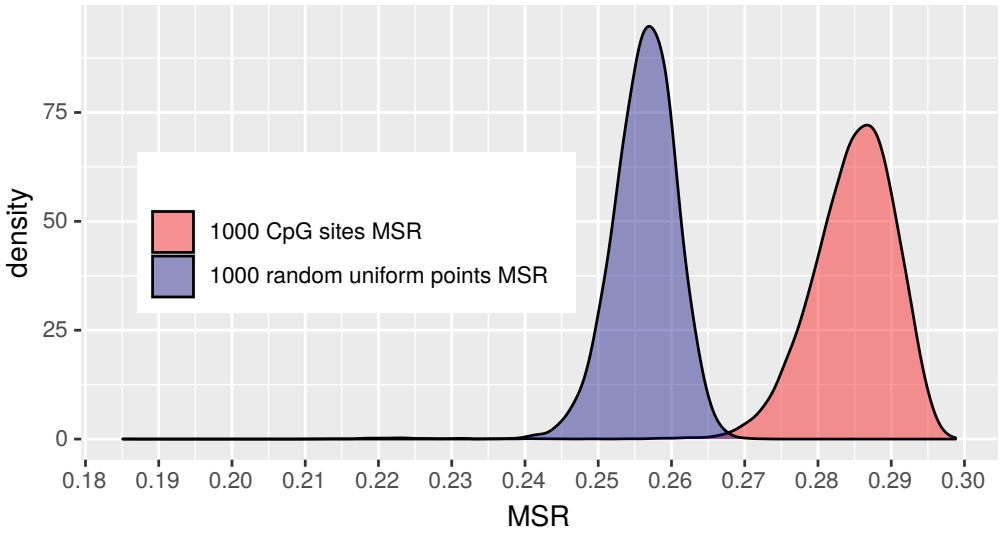

Fig. 3.12 The blue density curve was created calculating the MSR for 10^4 samples of 1000 points from a uniform distribution, the red one is based on $\approx 27,000$ fragments of 1000 consecutive CpG sites across the human genome. Here the MSR was calculated on the set of genomic position of each fragment of 1000 consecutive CpGs (irrespective of the methylation state).

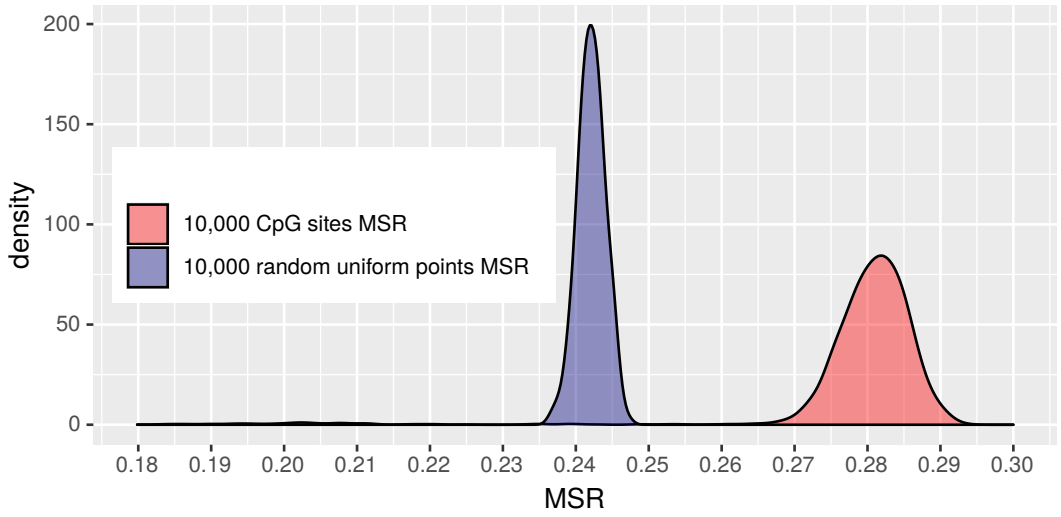

Fig. 3.13 The blue density curve was created calculating the MSR for 1000 samples of 10,000 points from a uniform distribution, the red one is based on ≈ 2700 fragments of 10,000 consecutive CpG sites across the human genome. Here the MSR was calculated on the set of genomic position of each fragment of 10,000 consecutive CpGs (irrespective of the methylation state).

3.7 Discussion

We saw there are genomic regions in which MSR_1 or MSR_0 assumes a significant value, most times a much smaller value than expected. To gain insight on the motivation behind this we can plot the methylation patterns of these particular regions, with the aim to get an intuition on how fragments look like in relation to MSR (and the other related statistics).

In fig 3.14 are shown 3 regions of 1000 CpGs with different $ecdf_0$. What is evident is that the fragment with low $ecdf_0$ shows a highly regular structure with a neat separation between methylated and unmethylated regions, while regions with a medium or high $ecdf_0$ have more "noisy" patterns. In general this holds also for the other fragments in the dataset.

An explanation can be the presence of unmethylated CpG islands that consists in areas of several consecutive unmethylated CpGs. Islands that are included in regions with a medium $ecdf$ instead have a less "clean" methylation pattern. Then regions with a density of CpGs around the mean value[5] ($\approx 1\%$) never have a very low $ecdf$ value, this can explain the positive correlation between the number of nucleotides and the significance measures.

The regularity in methylation patterns causes the *MSR* to drop, since the presence of long homogeneous regions makes a large fraction of bins, especially the small ones, to approach the maximum number of contained values k allowed by the bin size Δt, and then $H[k]$ to decrease. This behavior is a consequence of the application of MSR on data that is not generated by sampling values in an independent way, since when a certain value is sampled then it can't be sampled again. In fact we are able to obtain lower MSR values than what we expect to obtain with random independent samples from the most "unstructured" distribution, the uniform.

So in this case the small values of MSR are caused mainly by the high regularity at the smallest scales as illustrated in an example in 3.16.

This suggest that in this context one factor that makes *MSR* small is the auto-similarity in the methylation between close CpGs. In particular we expect that in these regions the proportion of methylated reads for a certain CpG is correlated to that of the next one. So we propose to characterize the fragments also by the auto-correlation between the methylation level of consecutive CpGs. As we can see from (3.17) that shows the correlations between region features and this new variable, the methylation autocorrelation has a clear relationship with the other significance measures. An advantage of calculating the autocorrelation is that it also takes into account the proportion of methylated reads of each site, beside the fact that it's easy to compute.

[5] sometimes called the "sea" as opposed to the CpG islands

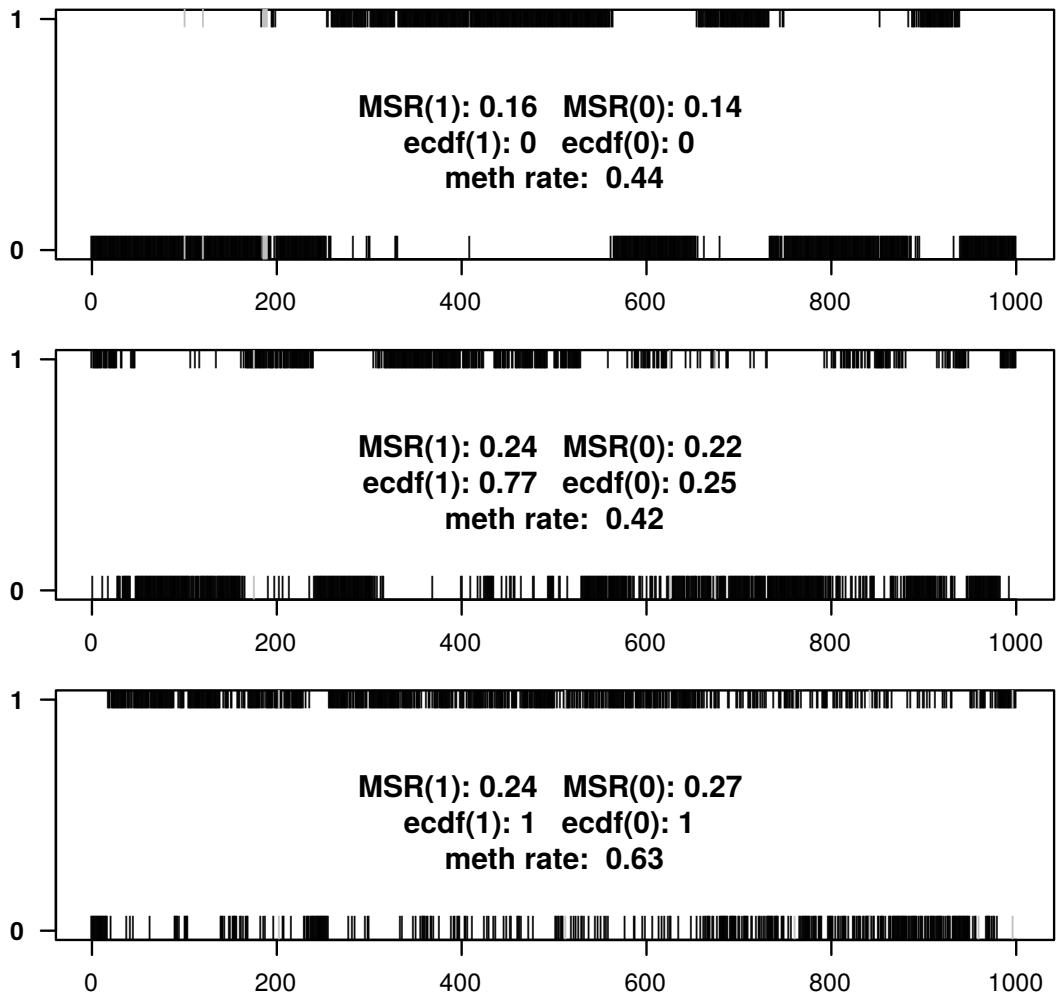

Fig. 3.14 In these plots are represented three binary strings to which MSR is applied, derived from three fragments of 1000 consecutive CpGs of stomach WGBS data. The fragments are ordered by increasing $ecdf_0/ecdf_1$.

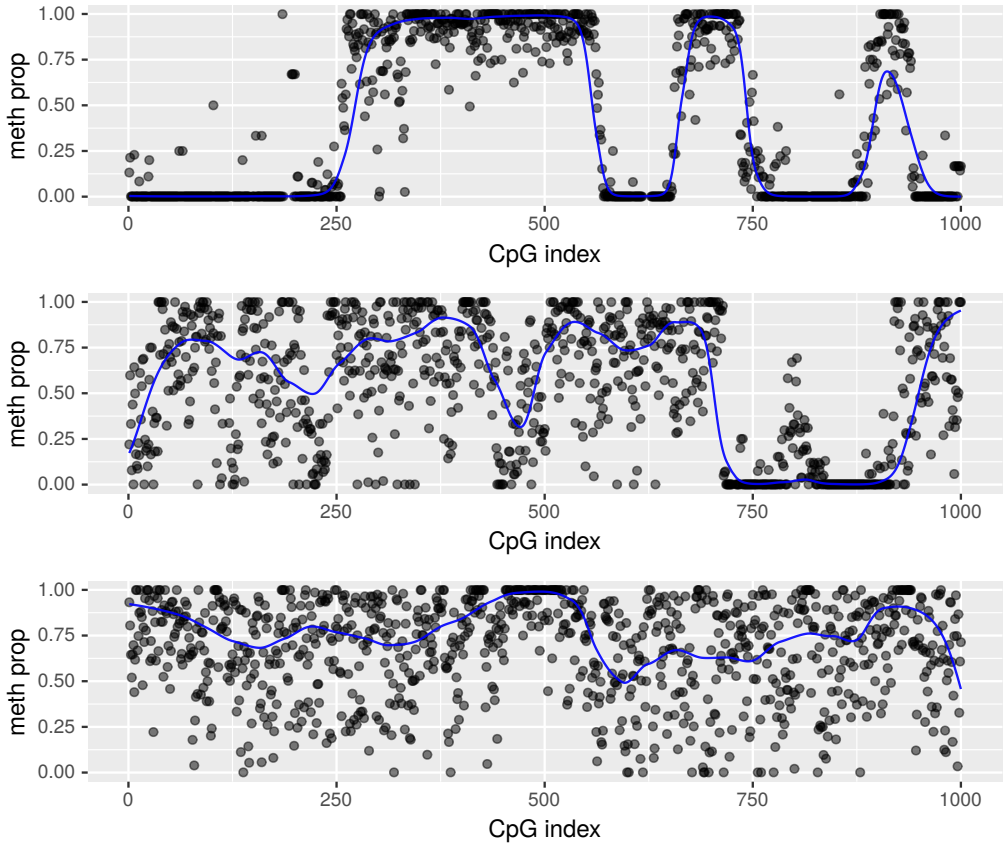

Fig. 3.15 These are the same regions shown in 3.14, but here points represent the proportion of methylated reads for each CpG.

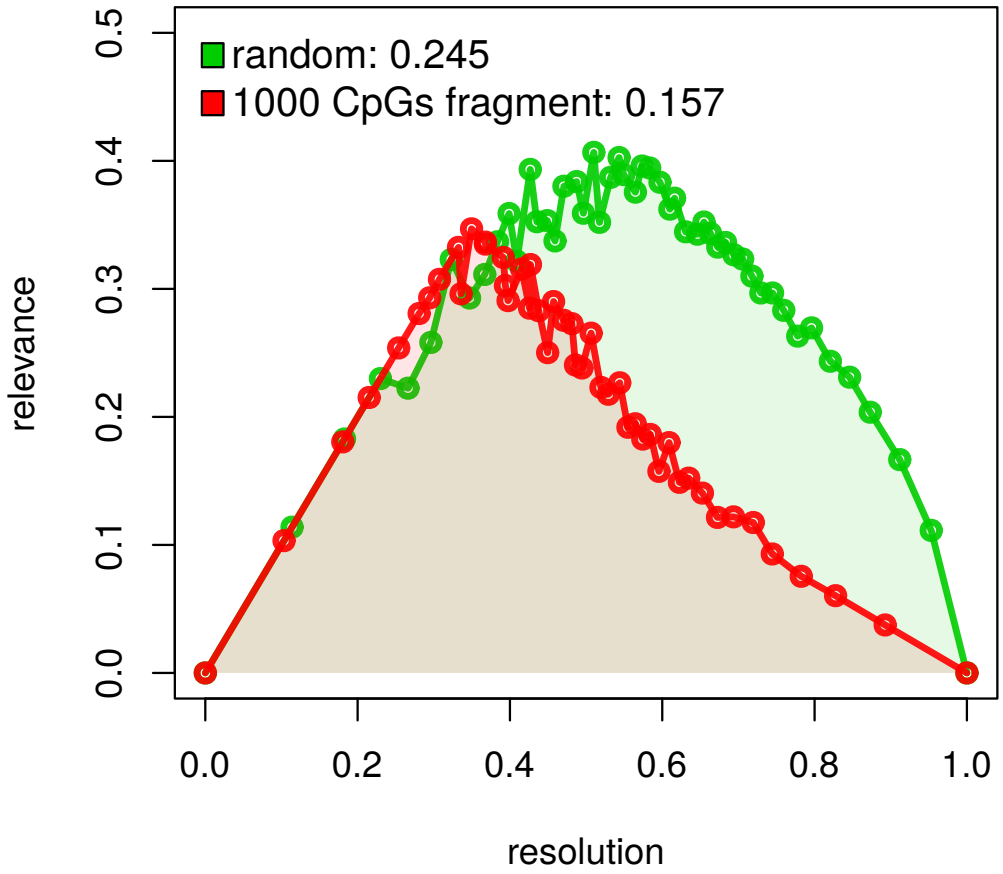

Fig. 3.16 The MSR curve was calculated for a fragment of 1000 CpGs characterized by an $ecdf_0 = ecdf_1 = 0$ and a methylation rate of ≈ 0.4. The other was based on a random binary string with the same length of the fragment and approximately the same proportion of ones. The first curve has lower values of $H[k]$ especially for high $H[s]$ values.

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

Fig. 3.17 *Pearson correlation coefficient between methylation autocorrelation and other features.*

Chapter 4

MSR and Expression relationship

We have seen how MSR and the related statistics can give a measure of the regularity of the methylation pattern that characterize a certain genomic region. We want now to explore the relationship between these features and the transcriptional activity that concern a certain genomic region. It was already discussed in the first chapter how some methylation patterns are believed to influence gene expression. The most evident mechanism is gene suppression through methylation of the promoter region, but some research suggest that a role is played also by methylation of islands' neighborhood, "the shores".

It was shown how MSR was employed to characterize datasets by measuring the information they contain on the hidden generative process, or from a different point of view, by highlighting the *relevance* of the observed variable with respect to the process that generated the data. The idea from the beginning was to find a covariate to be linked to the methylation pattern and in particular to MSR, the choice of expression is probably a trivial one, but it's also possible to investigate the relation of MSR with other genomic features.

4.1 Expression data

In a cell it's possible to find a multitude of distinct RNA molecules, with a large variety of functions. The set of all RNA transcripts, both coding and non-coding[1], in a population of cells is often called *transcriptome*[2]. The transcriptome is highly characterizing for cells of the same type or tissue, and defines their function in the organism. The gene expression machinery, that as we discussed is influenced by epigenetic modifications, determines how much each gene is expressed, so it is responsible of the observed transcriptome.

RNA-seq is a sequencing technique that aims to access this information, in particular it permits the estimation of how much a gene is expressed in relation to the others. So it does

[1] an RNA molecule can encodes for a protein or not

[2] The term depending on the context is both used to refer to all RNAs or just mRNA

not offer an absolute value of the quantity of each RNA transcript, instead it tells which fraction of the total amount was represented by a certain transcript.

There are mainly two methods in which RNA-seq can be performed, that differs in the way a type of RNA called ribosomal RNA (rRNA) is removed from the count of transcripts. The removal of the highly abundant rRNA is performed for the sake of an efficient transcript/gene detection of the other more relevant RNAs. The first is polyA+ selection, in which mainly RNAs with a molecular component named poly(A) tail (a long sequence of adenines) are selected. Almost all messenger RNA (mRNA) have a poly(A) tail, but it is not a characteristic of all RNAs molecules, so this technique ends up in focusing mainly on the protein-coding fraction of a transcriptome. The second is rRNA depletion (or total RNA-seq), in which ribosomal RNA is depleted, keeping in this way all the other kinds of RNAs. Although rRNA depletion offers more complete information on the transcriptome, polyA+ it's often preferred since it's cheaper and most studies focus just on the expression of protein coding genes. RNA-seq with these two techniques are called respectively polyA plus RNA-seq and total RNA-seq (Zhao et al. [24]).

RNA-seq data from ENCODE is found as a table in which each row contains an array of expression measures for a certain gene. For human RNA-seq data the number of rows of these tables is around $60,0000$[3], while we said that the number of genes in the human genome is approximately $20,000$. One of the reasons is that those tables take into account all genomic regions that can express a transcript besides the more "classical" set of $\approx 20,000$ protein coding genes.

In our work we will use pme TPM (posterior mean Transcript Per Million) as a measure of relative transcript abundance of a certain gene. The TPM for a certain gene t can be read as "for every 1,000,000 RNA molecules in the RNA-seq sample, TPM come from t". The fact that the measure is a posterior mean is due to the uncertain nature of the the mapping process of the reads to a certain genomic region.

There are several experiments in ENCODE for which both WGBS and RNA-seq data are produced from the same sample of cells. In particular in this study we choose a set of 8 cell types/tissues for which both data of an acceptable quality was available:

- **H1**: Human male embryonic stem cell line, also called H1-hESC.
- **K562**: Immortalised myelogenous leukemia cell line from a 53-year-old woman in 1970.
- **GM12878**: Lymphoblastoid cell line, produced from the blood of a female donor with northern and western European ancestry.
- **GM23248**: Fibroblasts taken from a skin punch of the arm of a 55-year-old man.

[3]This happens both for total RNA-seq and polyA RNA-seq. Therefore in polyA RNA-seq a large number of rows is associated to a null value of expression

- **HeLa-S3**: Line derived from cervical cancer cells taken in 1951 from a 31-year-old African-American woman.
- **Endodermal cell**: Endodermal cells from a healthy male originated from HUES64[4][5].
- **Lung**: Lung tissue cells from a 30-year-old woman.
- **Stomach**: Stomach tissue cells from a 54-year-old man.

We will use WGBS data and polyA plus RNA-seq data, with the exception of stomach for which only total RNA-seq data was available. The data was based on the hg38 version of human reference genome.

4.2 Genomewide relation between MSR and expression level

In the previous chapter we studied the application of MSR to the entire genome without focusing on any particular area, the idea is to do the same adding the covariate of expression. We will subdivide the genome in regions comprehending an equal number of consecutive CpGs, then we will associate to each region a value of expression. The value of expression is the sum of the expression values of the included genes, in particular we consider enough to include the transcription start site (TSS) of a gene[6]. Information about the genomic location of each gene can also be retrieved from ENCODE.

4.2.1 Dataset

For each of the cell types listed above the methylome was divided in fragments of 1000 CpGs discarding the ones having more than 10% of missing data or a number of nucleotides greater than $3.5x10^5$ [7]. Moreover since the cells come from different genders, sex chromosomes were discarded.

For each fragment we will indicate as $\boldsymbol{p}$ the array in which $\boldsymbol{p}_i$ is the proportion of methylated reads of the ith CpG.

Each fragment is characterized by an array of features that are divided in the three groups:

- **Basic features**

[4] HUES64 are human embryonic stem cell line derived from blastocysts

[5] Data is available on Encode but RNA-seq data was due to Roadmap Epigenomics

[6] Considering the full body would cause some ambiguity, since it can happen that certain genes are not fully included in a genomic region.

[7] These were just few outliers

  - **methylation rate**: mean of $\boldsymbol{p}$.
  - **nucleotides**: number of nucleotides included from the location of the first CpG to the last. Often it can be useful to consider the logarithm of this quantity instead.
  - **CpG density**: rate between the number of CpGs of the fragment and the number of nucleotides

- **Advanced features**
  - **methylation autocorrelation**: Pearson correlation coefficient between $\boldsymbol{p}_{i=1,\dots,l-1}$ and $\boldsymbol{p}_{i=2,\dots,l}$, with l the number of CpGs of the fragment.
  - **mean entropy**: the mean of the vector of components $-\boldsymbol{p}_i \log_2(\boldsymbol{p}_i) - (1-\boldsymbol{p}_i)\log_2(1-\boldsymbol{p}_i)$ for $i = 1,\dots,l$. It's the entropy of a Bernoulli distributed variable and here it is used as a measure of heterogeneity in methylation between the cells in the sample relative to a certain region. This is due to the fact that an high level of entropy is associated to values of $\boldsymbol{p}_i$ far from 0 or 1, that happens when the methylation state for a CpG is different between cells.
  - **methylation standard deviation**: standard deviation of vector $\boldsymbol{p}$.

- **MSR related features**: MSR_0, MSR_1, $residual_0$, $residual_1$, $ecdf_0$, $ecdf_1$ and $CpGsitesMSR$. This last feature indicates the MSR calculated for the genomic positions of the fragment's CpG sites, it's the same for all tissues since it doesn't depend on methylation.

For measuring expression log(pme TPM)[8] is used instead of pme TPM, it is often done in quantitative studies of gene expression as the log transformation is "better" distributed.

4.2.2 Analysis

Overall methylation levels may vary significantly from cell types, for example many cancer cell manifest a drop (Figure 4.1).

Despite this, cells are similar from the point of view of expression, their fragments' log(pmeTPM) are correlated with Pearson's correlation coefficients that ranges from 0.8 to 0.9, and their distributions have a similar shape (Figure 4.2).

Some features are highly correlated between them, as already said this is the case of $residual_0$, $residual_1$, $ecdf_0$, $ecdf_1$, while the correlation of these with MSR_0 or MSR_1 depends on the specific cell type. Between the advanced features methylation standard deviation is well correlated with autocorrelation (often $r > 0.7$), this is probably due to the fact that an higher standard deviation coincides with a smaller mean entropy and a

[8] More precisely log(pme TPM+ ε) where ε is a small valued constant in order to avoid the application of the logarithm to 0.

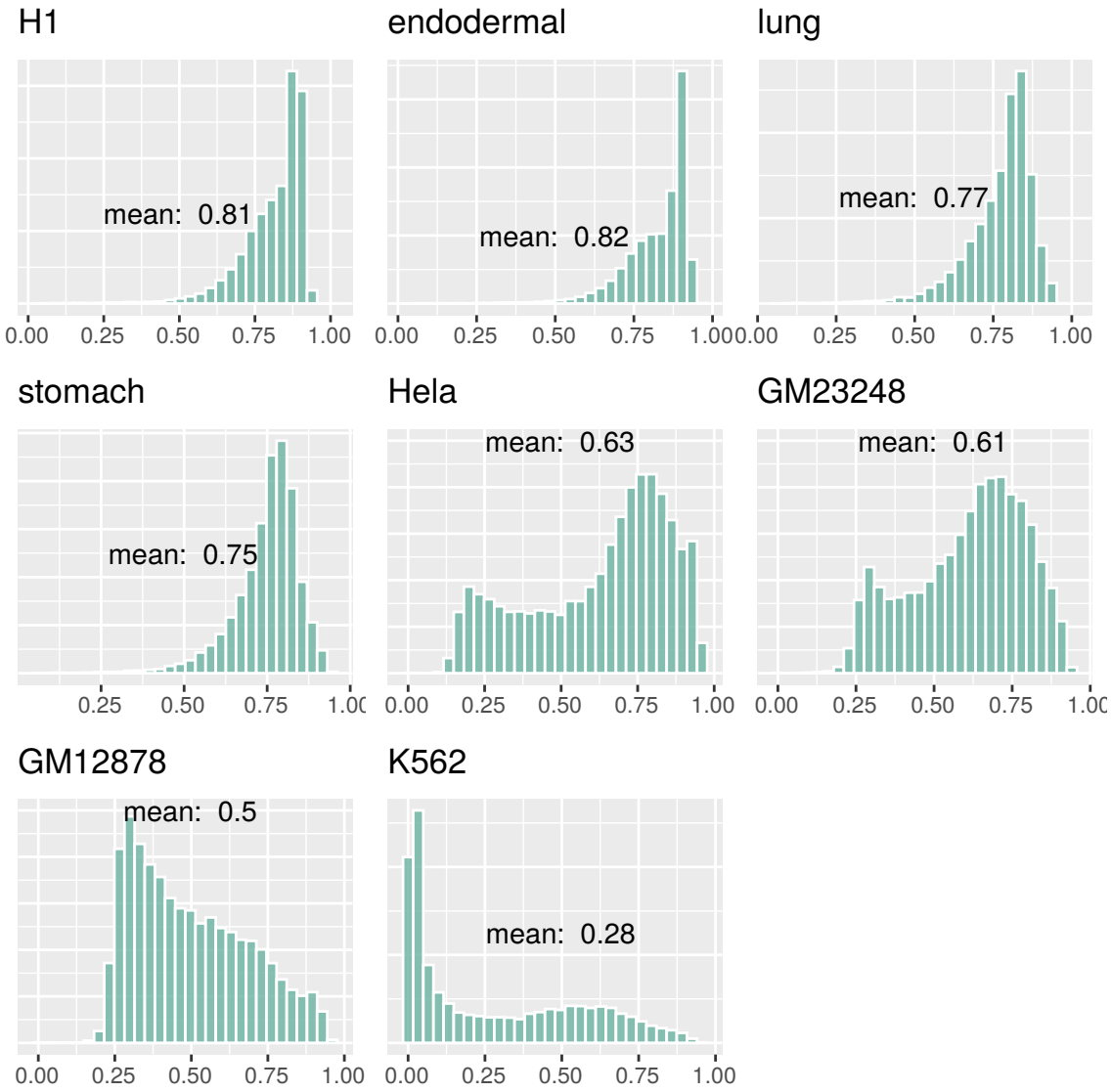

Fig. 4.1 *Fragments' overall methylation rate for several tissues.*

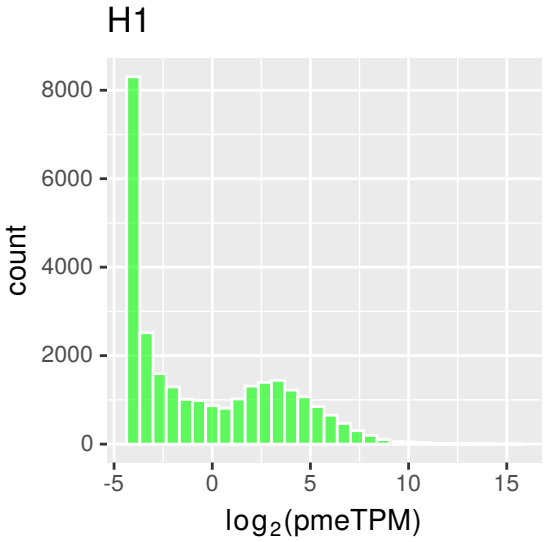

Fig. 4.2 *Distribution of* $\log_2$*(pmeTPM) in H1 fragments*. The peak at -4 is due to the default value assigned to regions having a null pmeTPM. The distribution is similar for all the cell types listed before. The bimodal shape suggests the presence of two different groups of fragments.

general methylation rate closer to 0.5, that in this context corresponds to regions having a well defined methylation profile. Then as a consequence of what it was discussed in the previous chapter these features result often correlated and in a certain sense "telling" us almost the same thing about a fragment.

Now we can start investigating the relationship with expression analysing the correlations between the features and the expression levels (Figure 4.3). The meth autocorrelation is the feature that often achieves the best r, which is often close to the value of one of the MSR related statistics (Figure 4.5). An interpretation can be that the autocorrelation characterizes functional structures in the genome, and moreover the more "clean" and "conserved" the better for expression.

The meth rate has a significant negative correlation in half of the cases, this is probably due to the fact that in these healthy cells with a normal overall methylation rate it could detect the presence of a unmethylated island, and exclude the "sea" regions. This was not true for the other cells, that have a lower overall methylation rate and more variability in the methylation rate of the fragments. Probably in those cases there are also "sea" or non-functional regions that encountered a considerable demethylation process. For example in K562 a higher methylation rate probably is a sign of a more "conserved" region.

Then the mean entropy r suggests that heterogeneity between cells is linked with a lower expression level, maybe because homogeneity is a trait of functional regions or

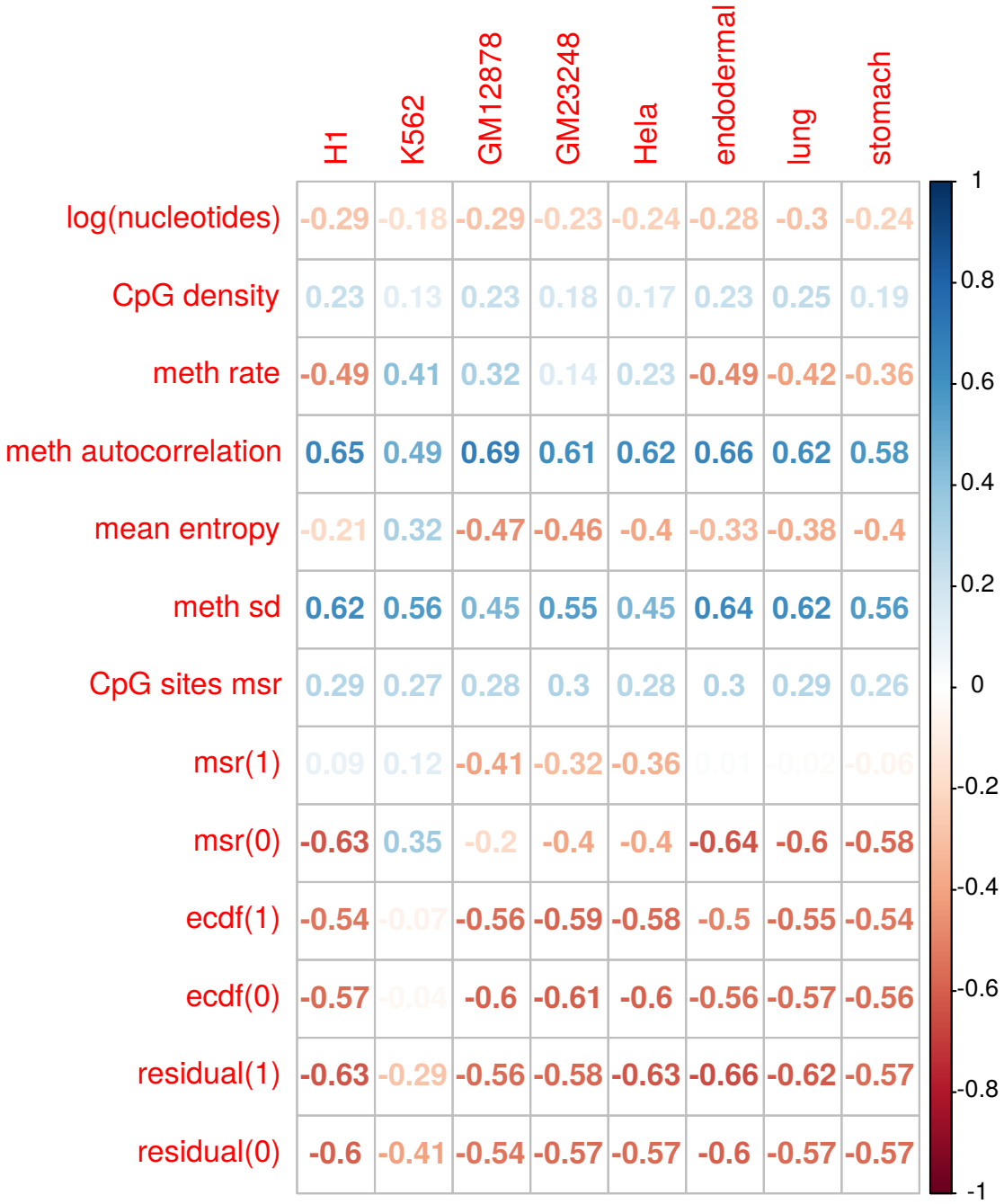

Fig. 4.3 *Correlation of several fragments' features with* log*(pmeTPM) for different cell types.* In this matrix each entry is the Pearson's *r* between a certain feature (row) with log(pmeTPM) for the fragments of a certain cell type (column).

because it highlights that ones in which cells with useful methylation patterns are facing a process of drift.

About MSR related statistics we can see that when MSR_0 or MSR_1 are lower than expected, so when there is a regular structure, the expression is higher, drawing in this way the same conclusion about autocorrelation. It is interesting to see this from a *meth rate-*MSR_0 scatterplot (Figure 4.4) where the interaction of these two variables highlights a relation with the expression level.

Finally we can observe that the CpG sites MSR has a small but still significant correlation with the expression (Figure 4.6).

4.2.3 Models

It was shown how methylation features alone were related to expression, the questions at this point are if it is possible to combine this features to model expression, which is their relative importance, and how much predictive can be a model based on these features.

The easiest possible model is the linear one. To see how much complex variables can improve the prediction, three different models are built adding each time one of the three groups of features listed above. To evaluate the performance a test set of 25% of the dataset is used, the rest is used as a train set. For each dataset a different model is built and then evaluated on the test set. Since we are also interested in a general model independent on the cell type, we build also a "shared" model merging the train datasets, that is then evaluated on the single test sets.

The linear models based on basic features achieve relatively poor performances, that significantly drop when generalizing in a single model for all cells. In fact methylation rate alone was not related to expression in the same way between cell types.

When adding the second group of predictors performances get significantly better, some of them are near to explain half of the variance of the data, moreover when building a model shared between cells types the results are just slightly worse.

Finally adding the MSR related features doesn't give a significant increase, this can be due to the high correlation between them, it seems that in this context they capture the same characteristic about fragments.

Looking at the estimated coefficients of these models can be misleading due to the correlations between features. It's possible to resort to regularization techniques in order to shrink coefficients, in particular we use Lasso regression as L1 norm penalization can force coefficients towards 0 and then perform a variable selection. With Lasso regression

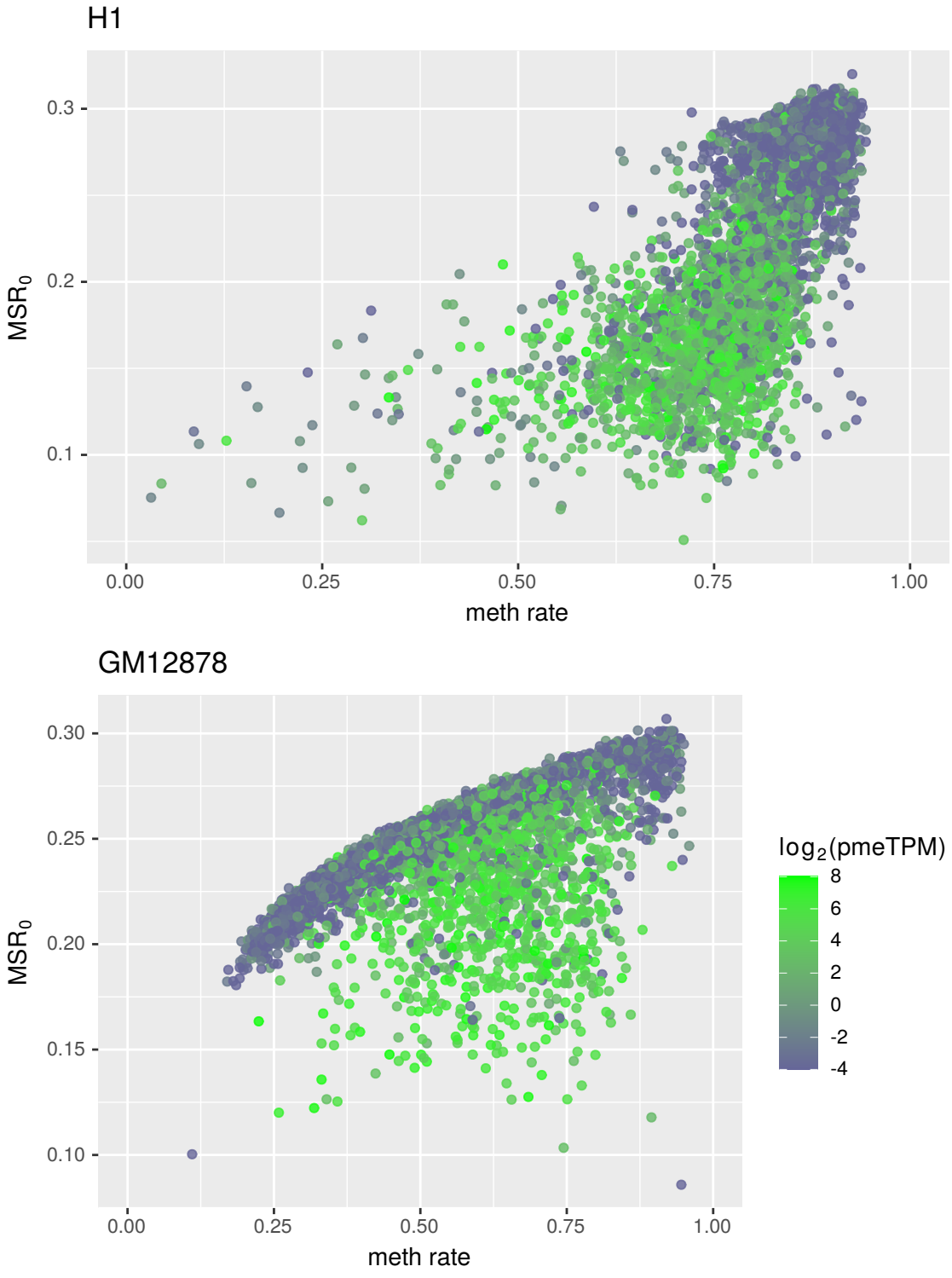

Fig. 4.4 In these plots each point represents a fragment (5000 were sampled from the dataset), and its color is the expression level associated to it. It's evident that when it has a broad distribution, methylation rate alone have a poor predictive power. Adding MSR_0 on the y axis highlights that points with an over-regular structure are more probable to have an higher expression level.

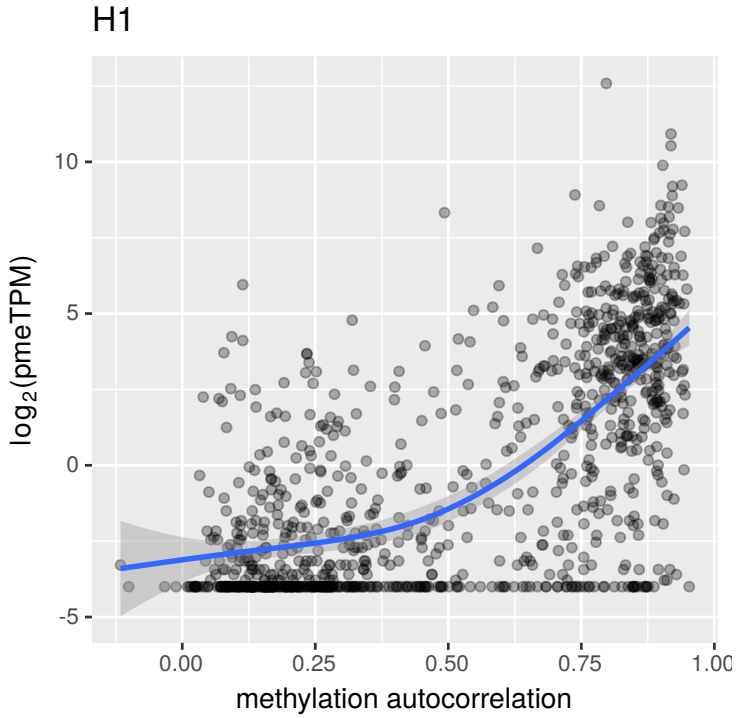

Fig. 4.5 *Methylation autocorrelation and expression level for a sample of fragments of H1 cells.*

we were able to obtain almost the same performances of the full shared model discarding same predictors (Figure 4.8).

The most important variables seems to be methylation standard deviation and methylation autocorrelation, then it is interesting to notice that the *CpGsitesMSR* was included in the model while the length in terms of nucleotides or the CpG density were discarded. The model was trained after standardizing the variables, so the absolute value of the estimated coefficients can give a measure of importance.

Finally we want to know how much predictive can be these features taking into account non-linear interactions between them. Next are shown results obtained training a model based on trees with gradient-boosting (Figure 4.9).

The improvement is moderate with respect to the best linear model but still relevant, moreover the algorithm also returns a measure of relative importance of the variables (Figure 4.10). Importance 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. [9]).

4.2.4 Discussion

This genome-wide analysis demonstrates that there is a relation between the methylation of a certain region and its expression beyond the overall methylation level. In particular fragments choice was unbiased with respect to the presence of genes and also to the pres-

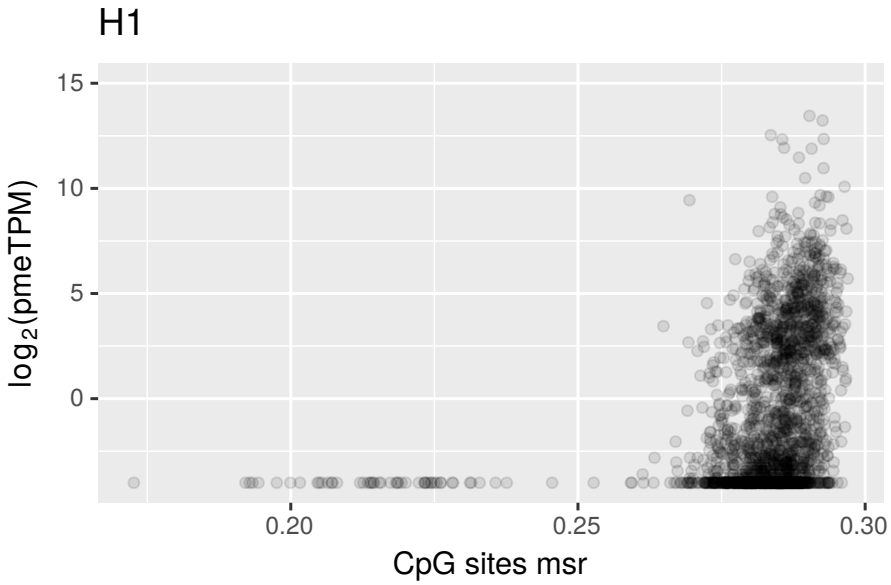

Fig. 4.6 The scatterplot shows a sample of 2000 fragments from the original dataset. We have already seen that we obtain high values of MSR (relatively to random sampled data) when calculating it on the positions of the CpG sites of a fragments. It is interesting to notice that there is a slightly positive correlation with expression, and points with a very low MSR have always a null value of expression. The meaning can be that MSR highlights regions were CpG sites have a non-trivial distribution that is the result of an optimization process, in this case we can think of the functional role in gene expression. Anyway the distribution is not broad and point with a low MSR are a very small fraction, maybe because "non functional" CpG groups tends to be less evolutionary conserved and then to disappear. This also shows the difference with MSR calculated using the CpG list (Figure 3.2) instead of genomic positions (Figure 3.4), in the first case it ends in giving a measure of spatial correlation, in the second case it's meaning is more related to the original one. This is due to the "improper" adaptation on binary strings, that is mitigated when the proportion of ones is small.

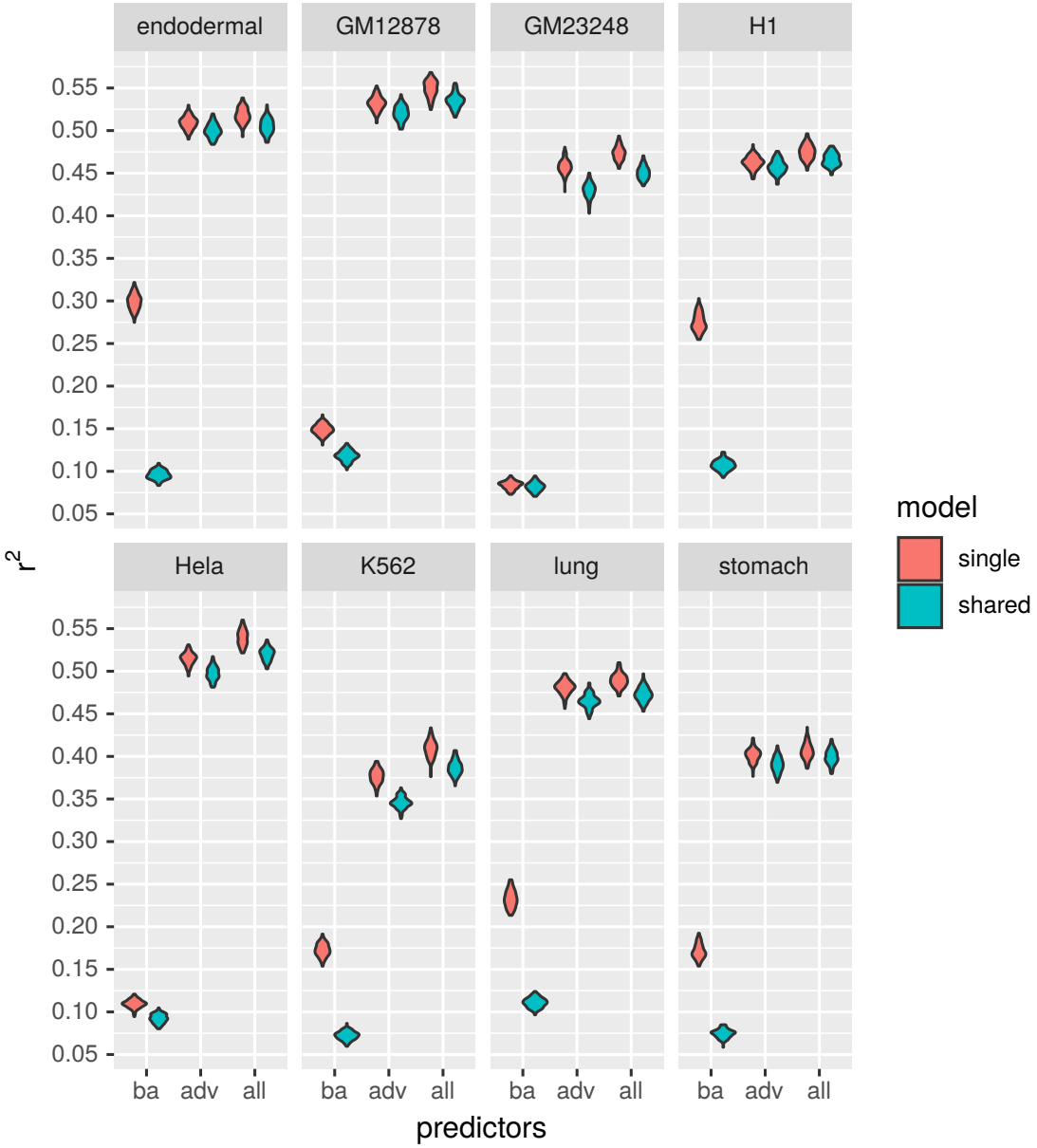

Fig. 4.7 *R squared coefficients for linear models with different sets of predictors.* The plot shows the performances obtained with linear models trained with different set of predictors (adding each time a new group) for different cell types. R^2 coefficient were calculated on a test set including 25% of the data for 100 random splits. The models were trained both on single cell datasets each time (red), and both building a model merging all the datasets (blue).

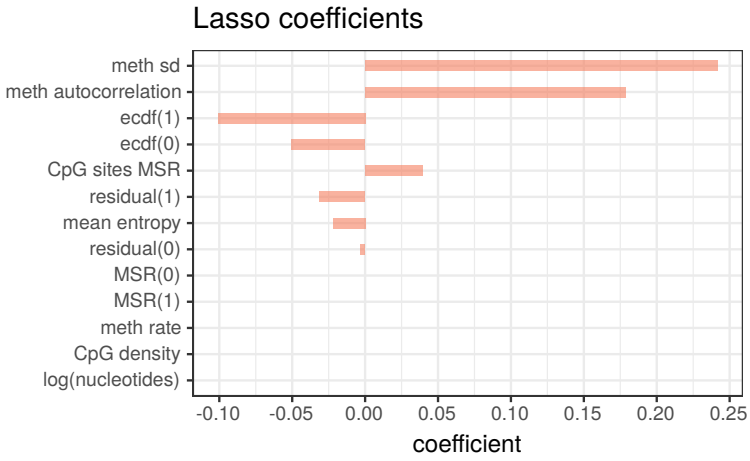

Fig. 4.8 Those coefficients were estimated training a Lasso model on data from all cells after standardizing fragments features (including expression level).

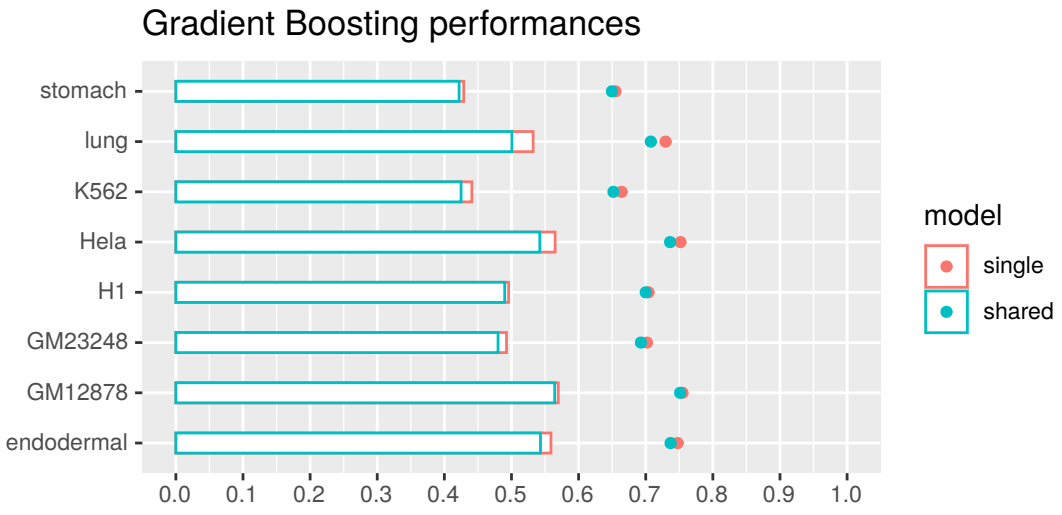

Fig. 4.9 A trees based gradient boosting model was trained on each dataset (red) and then one model was trained on a merged dataset (blue). Points are Pearson's r correlation coefficients of test sets with predictions, while horizontal bars are the r^2.

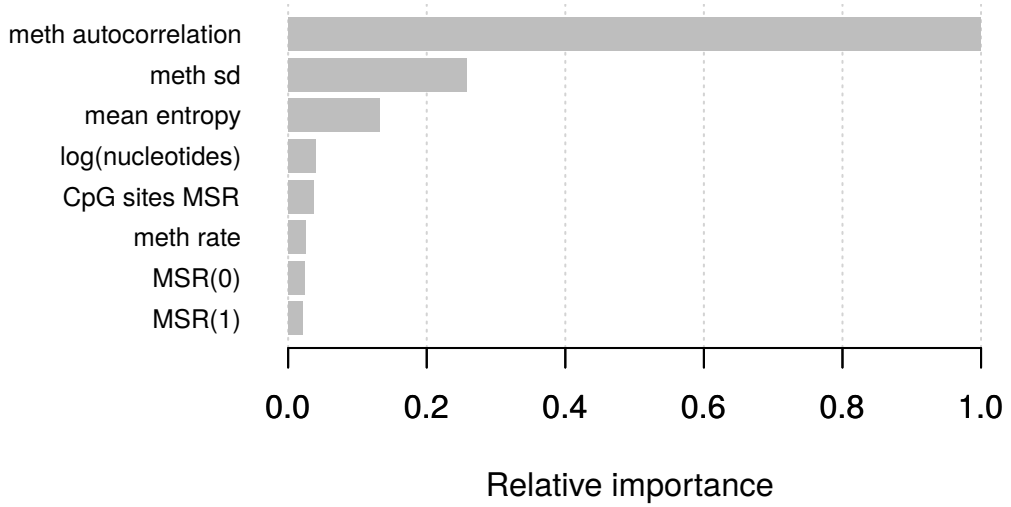

Fig. 4.10 *relative importance of variables used for the "shared" gradient boosting model.*

ence of known functional regions like promoters. Our interpretation is that the predictive power is due both to the ability to detect methylation functional structures like CpG islands and to measure their "health" state or "potential". Anyway this approach has limits, the division in fragments was arbitrary, while probably focusing properly a functional region could be needed, for example where a CpG island is just partially included in a fragment. In fact most studies concentrate in promoters and protein coding gene bodies.

Moreover we are considering just one of the multitude of epigenetic factor that are known to influence gene expression, which cause-effect relation with expression is not already clear.

Finally we report that repeating the analysis with fragments of 10,000 CpGs led to a moderate improvement (Figure 4.11).

4.3 Analysis at gene-level

The interest of the previous analysis was to investigate the relationship between methylation patterns and expression in regions defined by a set of CpG sites.

Now we will focus on genes, looking at the set of CpGs included in the gene body of $\approx 18,000$ protein coding genes. This allows us to decouple the task of predicting the expression of a gene from the one of predicting the presence of a gene in a certain region. This approach is closer to what is done in other studies and then results should be more comparable.

Mapping expression to the analysed regions is in this case trivial, each gene has its own measured TPM. What is less trivial is the choice of the genomic region around the genes to include in the analysis, since there are functional elements like promoters that are often not included in the gene body. For example in Kapourani and Sanguinetti [14]

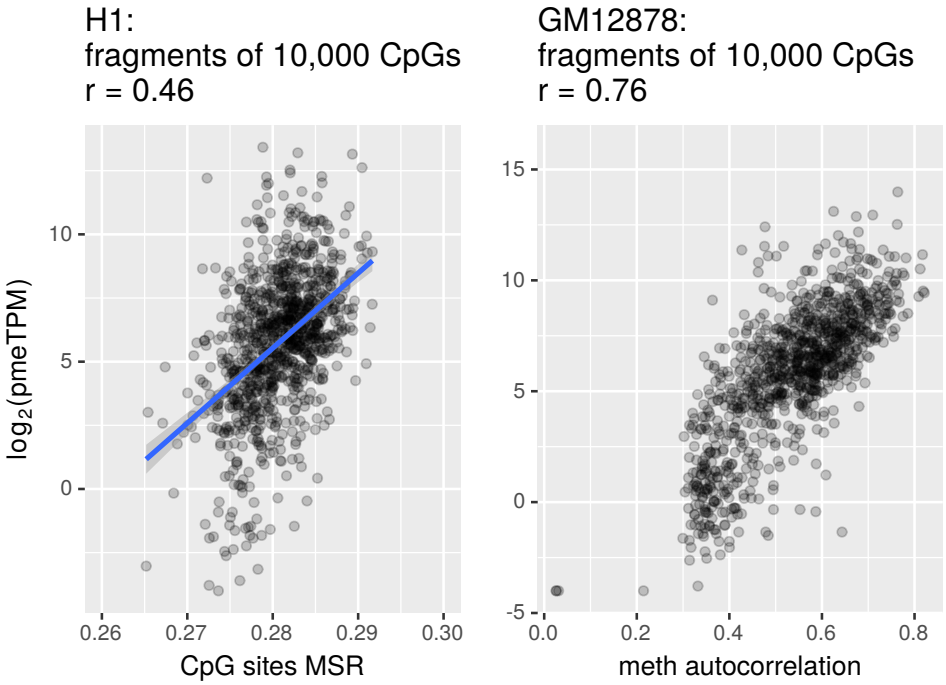

Fig. 4.11 Correlations with expression level are more marked when extending the observed window of CpGs to 10,000 (the two plots show a subsample of the entire dataset).

they choose to focus on the promoter region including then a window of a fixed number of nucleotides around the TSS. We will limit ourselves to gene bodies even though it would be interesting to enlarge the investigated area.

4.3.1 Dataset

The datasets are composed of a set of features describing the methylation of gene bodies for about 18,000 genes. We restrict the set of cells to six of them, discarding the two extracted from tissues (lung and stomach).

Since gene bodies don't include anymore a fixed number of CpGs we include it in the basic features.

Although even here MSR_1 and MSR_0 are often well correlated with expression, we chose to exclude MSR related features from the analysis since in several cases the number of CpG sites wasn't large enough to calculate them, and because of the correlation with other features.

As before only genes with less than 10% of missing data were kept, and sex chromosome were excluded.

4.3.2 Analysis

Overall methylation distributions are similar to what was observed for fragments, a difference is that demethylation on K562 is mitigated in gene bodies.

Also considering only genes, expression is correlated between different cells types with r around 0.8. The fact that we aren't dealing with regions with no genes, that are the same among different cells, made those correlations weaker, and moreover changed the distribution of expression (Figure 4.12).

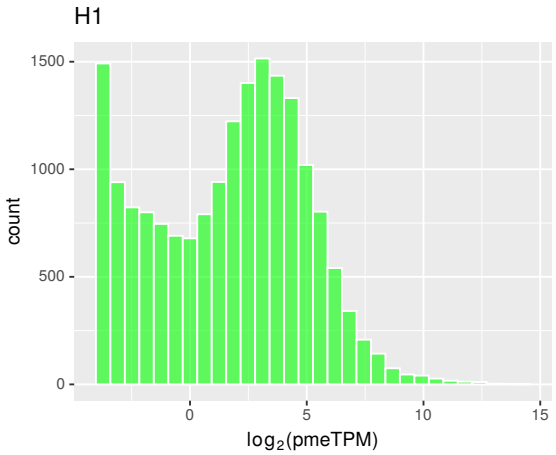

Fig. 4.12 *Distribution of* $\log_2$*(pmeTPM) for H1 protein coding genes.* The peak at the minimum is less marked with respect to the distribution of fragments since now only regions with genes are considered.

Looking at correlations of features with expression for the different cell types, methylation autocorrelation remains often the most correlated feature. The most evident difference is in the increasing of importance of the mean entropy. Methylation rate is slightly negatively correlated in H1 and endodermal cells, while is positively correlated in the rest of cells. The motivation could be that in healthy cells a low methylation rate is related to the presence of unmethylated islands, while in generally demethylated cells like K562 it denotes a region that encountered a process of drift.

4.3.3 Models

As done before for fragments we build linear models in a similar way, this time using two set of features. Another time we observe an improvement when including more complex features, in particular in the general model (shown in blue in Figure 4.15).

When estimating coefficients with Lasso regression, methylation autocorrelation and mean entropy seem to be the most important variables. (Figure 4.16).

We train a tree based model with gradient boosting to overcome the limits of linear regression. Figure 4.17 shows that we significantly improve performances for several cells,

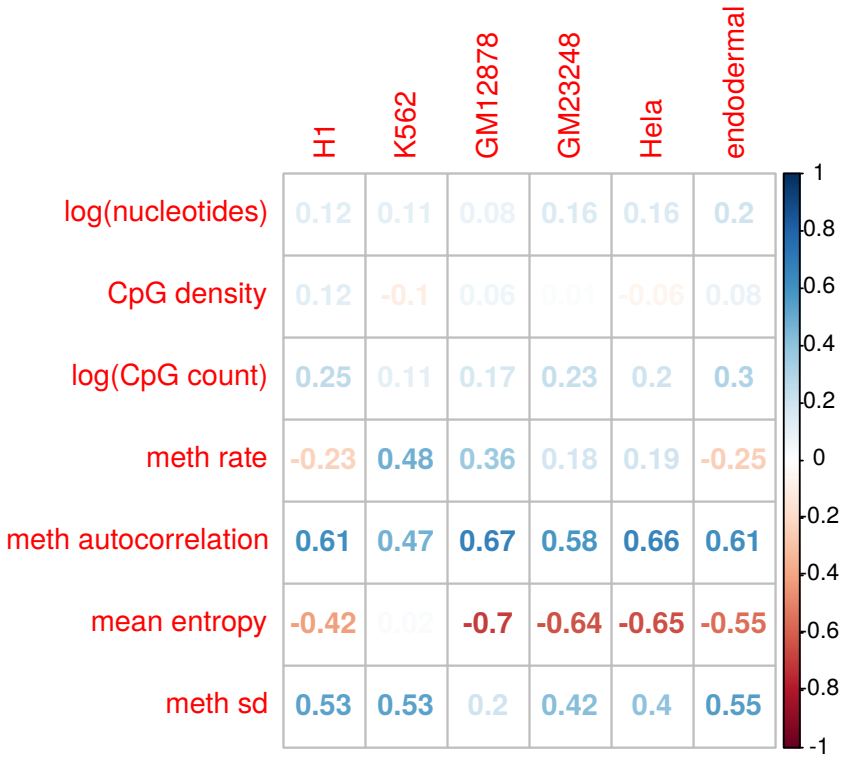

Fig. 4.13 *Correlation of several genes methylation features with* log*(pmeTPM) for different cell types.*

in particular for K562, that probably benefits from the model flexibility in dealing with highly different methylation patterns (with respect to other cells). The increment instead was moderated for H1 and the similar endodermal cells, confirming in this way that prediction of gene expression is hard for H1 genes. The gradient-boosting relative importance measure confirms methylation autocorrelation and entropy as the most important variables (Figure 4.18).

4.3.4 Discussion

Repeating the analysis focusing on genes demonstrates that methylation patterns not only are able to predict the presence of a gene (a region that could potentially express transcripts) but can also give an indication on how much a gene is active.

A question left unanswered is why performances are so different between cell types. A possible answer is that variance in expression is a bit higher in HeLa and GM12878 with respect to H1, and this part of variance can be well explained by the new predictors we introduced.

This is confirmed by the fact that the difference in log(pme TPM) between H1 and GM12878 is correlated with the difference in methylation autocorrelation and with mean entropy with an $|r| \approx 0.4$ for both features (in HeLa $|r| \approx 0.35$) while this is not true for

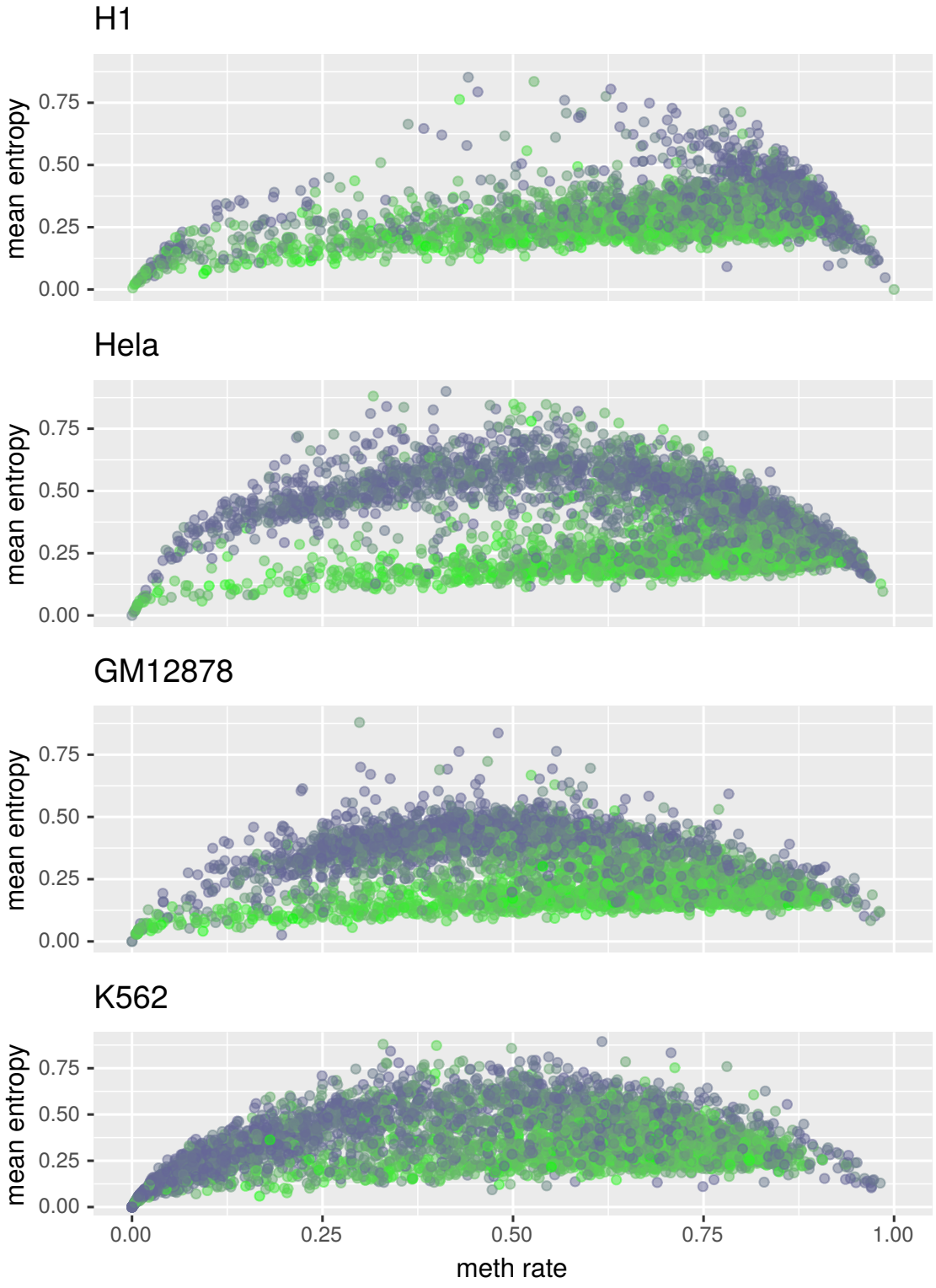

Fig. 4.14 In these plots each point represents a gene (5000 were sampled from the dataset), and its color is the expression level associated to it. As in Figure 4.4 the color indicates the level of expression (see its legend).
In general genes with a low mean entropy in methylation have a higher level of expression. Anyway in K562 this is not true, since there is a cluster of highly demethylated genes that have also a low entropy.
It is not clear if entropy itself causes a reduction in the expression level, of it is a mark of the occurrence of another phenomenon that causes a drop in expression.

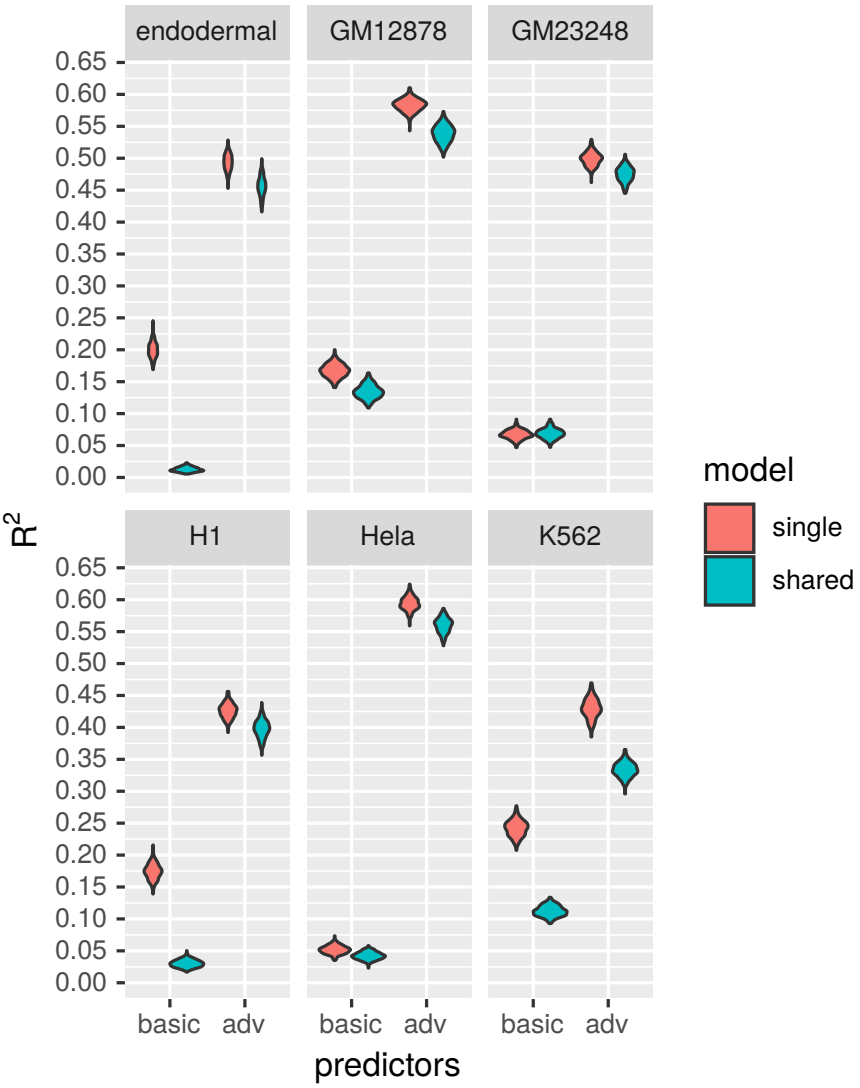

Fig. 4.15 *R squared coefficients for linear models with different sets of predictors.* The plot shows the performances obtained with linear models trained with two different sets of predictors (the first is the basic set, then advanced features are added) for different cell types. R^2 coefficient were calculated on a test set including 20% of the data for 300 random splits. The models were trained both on single cell datasets each time (red), and both building a model merging all the datasets (blue).

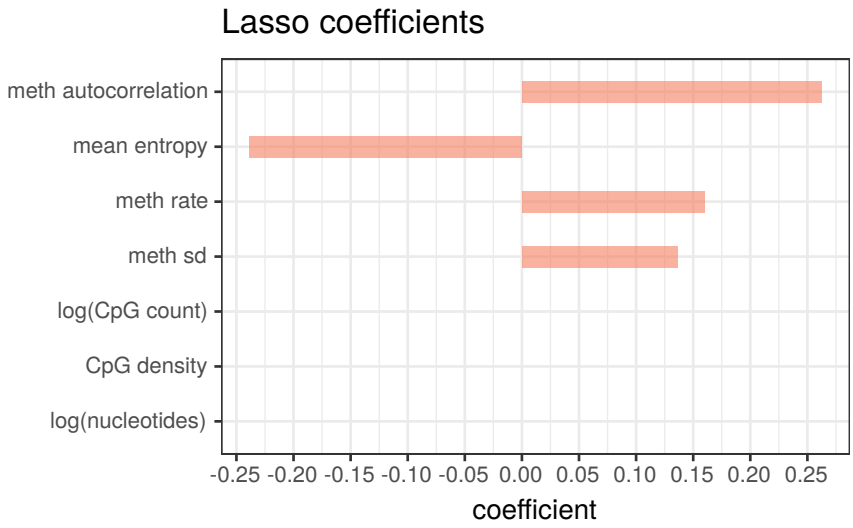

Fig. 4.16 Those coefficients were estimated training a Lasso model on data from all cells after standardizing genes' methylation features and expression level.

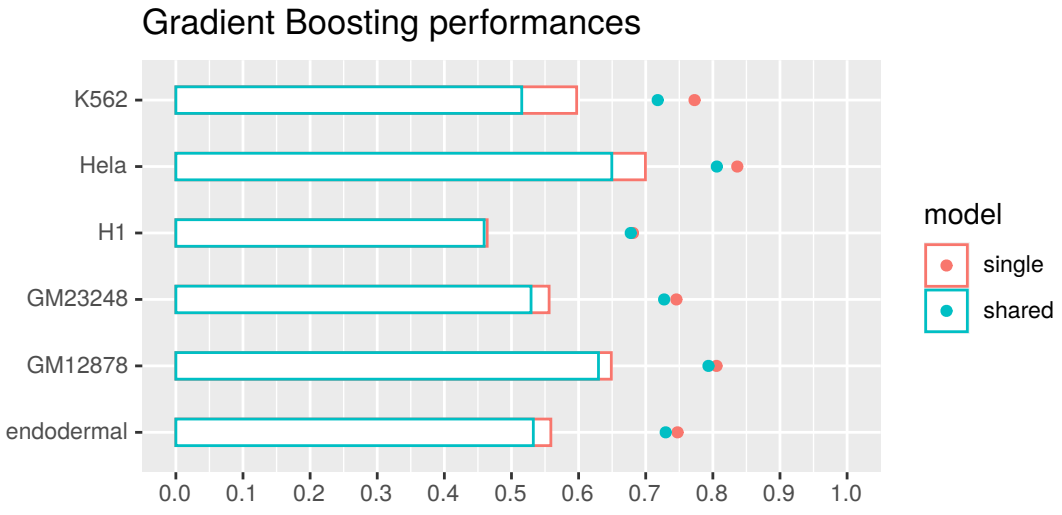

Fig. 4.17 A trees based gradient boosting model was trained on each dataset (red) and then one model was trained on a merged dataset (blue). Points are Pearson's r correlation coefficients of test sets with predictions, while horizontal bars are the r^2.

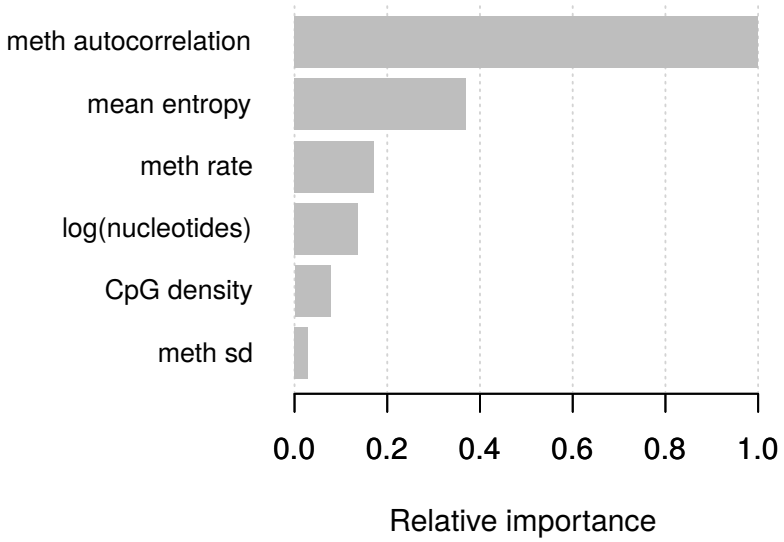

Fig. 4.18 *relative importance of variables used for the "shared" gradient boosting model.*

K562 in which the correlation is in mainly with difference in methylation sd with $|r| \approx 0.5$.

For GM23248 and endodermal cells instead it was not particularly correlated to any difference in features.

This investigation confirms the functional role of spatial correlations in methylation patterns, and also the utility of measuring the heterogeneity (for example with entropy) in methylation between cells. With respect to this, limiting the analysis to the general level of methylation gives poor insights, this underlines the importance of high resolution data in methylation (the possibility to know methylation state at CpGs resolution).

Anyway there are large margins of improvement, we can see from Figure 4.20 or Figure 4.19 that there are regions with a striking misleading expression level according to our models. In some cases we can suppose that the observed window is not large enough to include other functional sites, or maybe these features are too much generic to capture certain relevant behaviors, but in general we can think of unobserved covariates that influence expression.

The reported positive correlation of methylation in gene bodies with expression is consistent with these findings, since it's implied by an high autocorrelation, and a well defined separation between methylated and unmethylated areas. In general our results are qualitatively confirmed by other studies in which the methylation profile of functional regions was investigated, suggesting that "canyons" (Edgar et al. [8]),"ravines" (Jeong et al. [13]) or "U-shaped" (Kapourani and Sanguinetti [14]) regions are associated with an enhanced transcriptional activity. Main differences in this works are in the quantitative

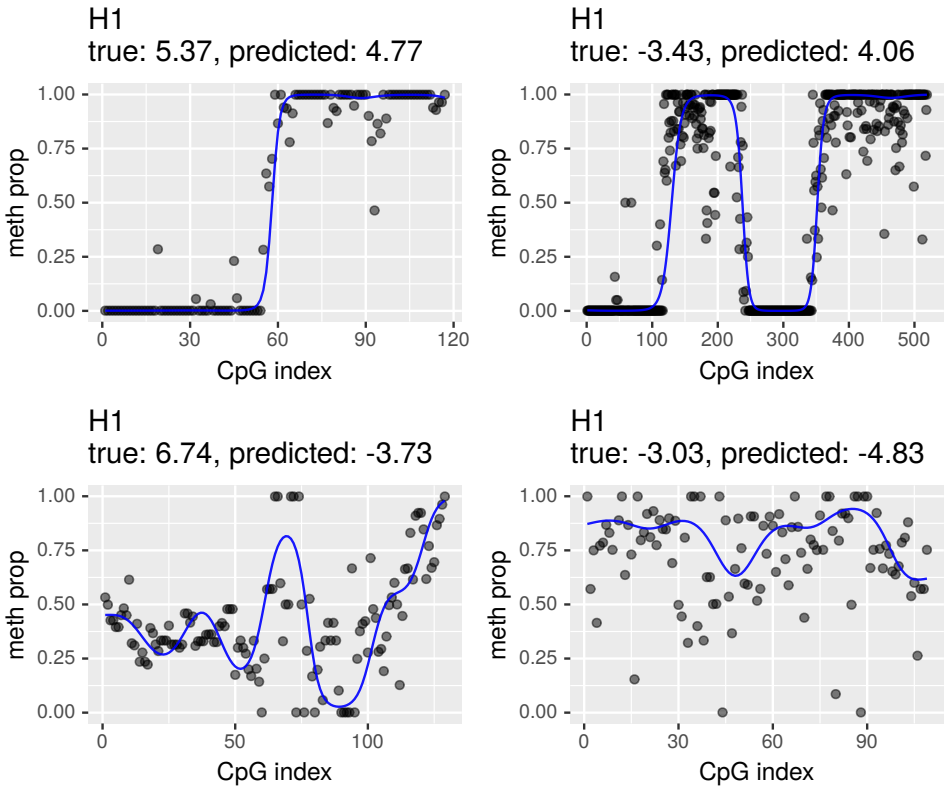

Fig. 4.19 *Four examples of genes' methylation pattern with true and predicted expression value.* The four genes were selected in order to show an example for each combination of low/high true/predicted level of expression. The model was linear and based on basic and advanced features.

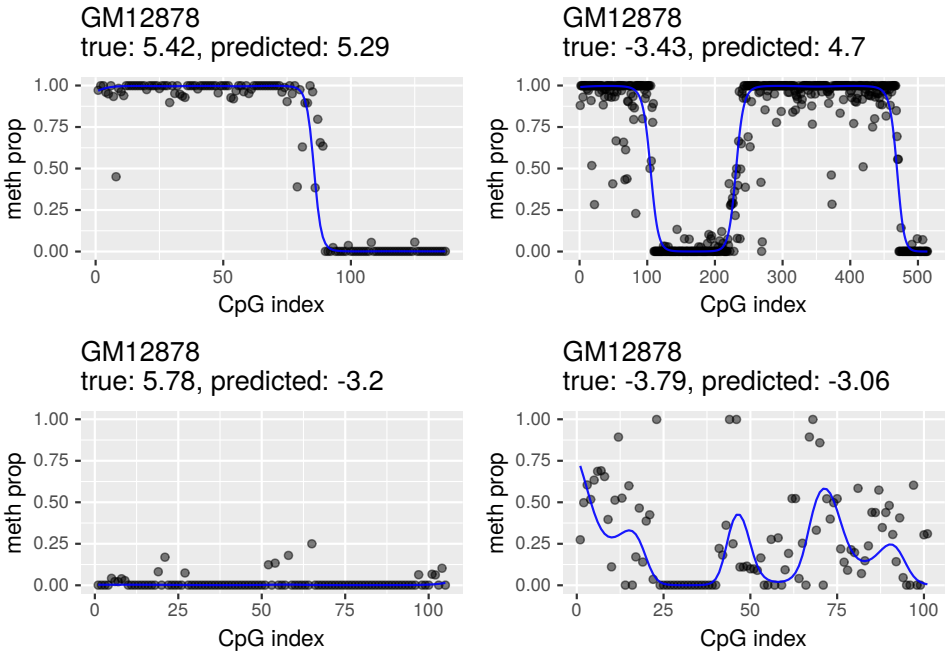

Fig. 4.20 See Figure 4.19 description

characterization of these methylation features, and in the broad genome-wide span of the analysis.

Discussion

The aim of this thesis was to use *multiscale relevance* to investigate a relation between the information that methylation patterns contain and gene expression.

After analysing the statistical behavior of MSR with random data, we applied it to different regions of the genome for a WGBS dataset. The intent was to study its behavior across all the genome and check if it could highlights certain regions. We applied MSR mainly in two in different ways: The first consists in considering only the methylation state of a list of consecutive CpGs irrespective of CpGs dinucleotides spatial distribution, the second instead consists in using the genomic positions of a group of contiguous CpGs.

The first method was the most explored, although being an adaptation probably distant from the original concept of MSR. The occurrence of low values of MSR highlighted in an unexpected way methylation patterns with a highly regular structure. These regions are characterized by a neat separation between methylated and unmethylated areas. So in this context MSR ended in detecting structures like CpG islands and quantifying their regularity. This is justified by the ability of MSR to quantify the richness of different densities across different spatial scales. Then we noticed that this characteristic detected by MSR could be quantified also with the autocorrelation in methylation levels, exploiting in this way also the "proportion" information of each site .

The second method was more similar to the original concept of MSR, and gave a totally different kind of information. It was in almost all cases higher than expected, showing in this way that the distribution of CpGs is non trivial. Anyway it was hard to give an interpretation to MSR in this context. At first glance it didn't highlight particular regions with recognizable characteristics, with the exception of a small fraction of regions with a very low value that show a uniform distribution of CpG sites. The distribution of MSR suggests that the spatial positioning of CpG dinucleotides is due to a non-trivial process, but anyway there is no evidence of a process of maximization of the information content in any of the observed regions.

Finally, we used this findings to characterize genomic areas besides other "classical" features and we related them with expression. Modelling expression utilizing only the mean methylation level and the density of CpGs gives in general poor results, moreover the model didn't generalize to several cell types. Adding the meth autocorrelation and a measure of

heterogeneity between cells methylation dramatically improved the prediction accuracy. The improvement is mainly due to the high correlation of methylation autocorrelation with expression. Adding features related to MSR instead has not led to significant improvements, despite the good correlation with expression similar to the one of autocorrelation.

From models, that were often able to explain about half of the variance, emerge that transcriptional activity is generally higher where i) methylation levels are spatially correlated ii) there is homogeneity between cells in methylation iii) the mean methylation level is not extremely low or high. This corresponds to regions that show the presence of both methylated and unmethylated areas that are neatly separated. These considerations generally hold both when considering a generic region in the genome of a fixed number of CpGs and when focusing on gene bodies.

These findings are consistent with recent studies that investigated the regulatory role of CpG shores and of the "shape" of the methylation profile. In our opinion the novelty and strength of our approach is in its generality, since the same analysis was done for a wide and unbiased choice of regions, and moreover for several datesets of different cell types. We remark that our approach was mainly directed at explaining the difference in expression level that characterizes different genes of the same cell, rather that comparing the same gene in a population of cells.

Although we think we gained some insights about the influence of methylation patterns on genes expression, the approach illustrated in this thesis work has several limits. Obviously the prediction accuracy was still poor, there were a multitude of regions for which the predicted expression value was far from the expected, both by our models and by our intuition after a visual inspection. Although we know there are other factors that influence expression, we could also think that we are missing some relevant methylation characteristics. A description in terms of summary statistics could be simplistic, we could consider other methylation traits. For example we ignored the extension, the number, and the position relative to the gene of the structures we detected. It's probable that a correct focus on a functional region is needed, for example when considering genes we could extend the attention on a larger area, in fact it can happen that regulatory elements like promoters are entirely outside the gene body. Another weakness was the handling of sites with a small number of reads, probably a correct approach would take into account the confidence of the measured methylation level between CpG sites.

An interesting perspective is the application of this method to single-cell data, to overcomes the limits of averaged data. Observing how expression level varies between the same sample of cells in relation to the methylation patterns could confirm those findings, especially where there is heterogeneity in methylation patterns. In general we could think of applying this approach for the same gene in a population of different cells.

Another future direction could be to study how the characteristics we found are related

to the tissue specificity of the genes, for example separating the genes in two groups: tissue-specific genes and house-keeping genes.

Finally, some considerations have to be done on the utility of *multiscale relevance*. MSR used in the first way described above resulted useful in detecting hyper-regular structures, but it wasn't itself more useful than other a posteriori extracted features in predicting expression. When used in the second way, it has a slight but interesting correlation with expression, but it was hard to interpret the meaning. At the end we have still to understand what MSR could tell us about the genomic spatial distribution of CpG sites. A disadvantage of MSR is in the dependence on the total number of points M, that becomes more complex in this adaptation on "discrete" data. Then it requires large samples to obtain significant measures, and it is more computationally demanding than simple statistics. Despite this, it has been proved useful for our purposes. Further experiments could involve the application of MSR on the genomic positions of methylated or unmethylated sites[9], and the analysis of the relationship with other covariates.

[9] We've already observed that in highly methylated areas MSR applied on all CpGs sites positions was highly correlated with MSR calculated only on the positions of the methylated ones. Anyway we haven't deepen it so much, since we mainly concentrated on MSR applied in the first way described above.

References

[1] Annunziato, A. (2008). Dna packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26.

[2] Centre for Genetics Education (2019). An introduction to dna, genes and chromosomes: https://www.genetics.edu.au/publications-and-resources/facts-sheets/fact-sheet-1-an-introduction-to-dna-genes-and-chromosomes. [Online; accessed 21-November-2020].

[3] Consortium, E. P. et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640.

[4] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

[5] Crick, F. H. (1958). On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8.

[6] Cubero, R. J., Jo, J., Marsili, M., Roudi, Y., and Song, J. (2019). Statistical criticality arises in most informative representations. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(6):063402.

[7] Cubero, R. J., Marsili, M., and Roudi, Y. (2020). Multiscale relevance and informative encoding in neuronal spike trains. *Journal of computational neuroscience*, 48(1):85–102.

[8] Edgar, R., Tan, P. P. C., Portales-Casamar, E., and Pavlidis, P. (2014). Meta-analysis of human methylomes reveals stably methylated sequences surrounding cpg islands associated with high gene expression. *Epigenetics & chromatin*, 7(1):28.

[9] Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813.

[10] Grigolon, S., Franz, S., and Marsili, M. (2016). Identifying relevant positions in proteins by critical variable selection. *Molecular BioSystems*, 12(7):2147–2158.

[11] Haimovici, A. and Marsili, M. (2015). Criticality of mostly informative samples: a bayesian model selection approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(10):P10013.

[12] Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):3156.

[13] Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., et al. (2014). Large conserved domains of low dna methylation maintained by dnmt3a. *Nature genetics*, 46(1):17–23.

[14] Kapourani, C.-A. and Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412.

[15] Mandal, Ananya (2019). What is dna? [Online; accessed November 21, 2020, https://www.news-medical.net/life-sciences/What-is-DNA.aspx].

[16] Marsili, M., Mastromatteo, I., and Roudi, Y. (2013). On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09003.

[17] Moore, L. D., Le, T., and Fan, G. (2013). Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38.

[18] National Center for Biotechnology Information (2020). Catalytic domain. [Online; accessed 21-November-2020].

[19] Song, J., Marsili, M., and Jo, J. (2018). Resolution and relevance trade-offs in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(12):123406.

[20] The Hardin lab (2020). http://worms.zoology.wisc.edu/zooweb/phelps/zwk99004k.jpeg. [Online; accessed November 21, 2020].

[21] Tost, J. (2018). *DNA Methylation Protocols*. Springer.

[22] University of Leicester (2020). Dna, genes and chromosomes: https://www2.le.ac.uk/projects/vgec/schoolsandcolleges/topics/dnageneschromosomes. [Online; accessed 21-November-2020].

[23] Wikipedia contributors (2020). Dna methylation — Wikipedia, the free encyclopedia. [Online; accessed 21-November-2020].

[24] Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and von Schack, D. (2018). Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rrna depletion. *Scientific reports*, 8(1):1–12.