“La presente tesi è stata prodotta durante la frequenza del corso di dottorato in Medicina Molecolare dell’Università degli Studi di Cagliari, XXIX ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività I.3.1 “Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell’ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali”.

Università degli Studi di Cagliari

PhD Programme

MOLECULAR MEDICINE (Cycle XXIX)

Department of Biomedical Sciences

Unit of Biochemistry, Biology and Genetics

TITLE

Early detection of colorectal cancer: biomarker discovery

Scientific Disciplinary Code

BIO/13

PhD student: Antonio Fadda

PhD Programme Coordinator: Prof. Amedeo Columbano

Tutor: Prof.ssa Patrizia Zavattari

Final exam 2015 – 2016
Thesis discussed on Aprile 2017

Table of contents

Introduction

Chromatin structure defines the state in which genetic information is organized within a cell. This organization of the genome into a precise compact structure greatly influences the abilities of genes to be activated or silenced. Epigenetics, originally defined by C.H.Waddington [1] as 'the causal interactions between genes and their products, which bring the phenotype into being', involves understanding chromatin structure and its impact on gene function. Waddington's definition initially referred to the role of epigenetics in embryonic development. However, the definition of epigenetics has evolved over time as it is implicated in a wide variety of biological processes. The current definition of epigenetics is 'the study of heritable changes in gene expression that occur independent of changes in the primary DNA sequence'. Most of these heritable changes are established during differentiation and are stably maintained through multiple cycles of cell division, enabling cells to have distinct identities while containing the same genetic information. This heritability of gene expression patterns is mediated by epigenetic modifications, which include methylation of cytosine bases in DNA, post-translational modifications of histone proteins as well as the positioning of nucleosomes along the DNA. The complement of these modifications, collectively referred to as the epigenome, provides a mechanism for cellular diversity by regulating what genetic information can be accessed by cellular machinery. Failure of the proper maintenance of heritable epigenetic marks can result in inappropriate activation or inhibition of various signaling pathways and lead to disease states such as cancer.

DNA methylation: an overview

DNA methylation is perhaps the most extensively studied epigenetic modification in mammals. It provides a stable gene silencing mechanism that plays an important role in regulating gene expression and chromatin architecture, in association with histone modifications and other chromatin associated proteins. In mammals, DNA methylation primarily occurs by donation of a methyl (CH3) group from S-adenosylmethionine to the fifth position in the cytosine pyrimidine ring resulting in the formation of 5–methylcytosine (5mC). The reaction is catalyzed by a family of enzymes called DNA methyltransferases (DNMTs). 5mC is predominantly found within the genome in the context of 5′-cytosine-

phosphateguanine-3′ (CpG) dinucleotides. CpG sites are globally rare throughout the genome [2] and predominantly (~85%) constitutively methylated in healthy cells [3]. Approximately 10% of CpG sites, are densely concentrated into "CpG islands" (CGIs) regions in the genome where the percentage of the CpG dinucleotides is higher than would be expected based upon a random distribution of nucleotides. Of interest, CpG sites are generally under-represented in the genome presumably due to the susceptibility of 5mC to undergo transition mutations secondary to deamination [4]. CpG islands are often defined as sequences greater than 200–500 bases in length with greater than 50% GC content and a ratio of observed to expected CpG ratio greater than 0.6 [5]. More than half of the CGIs are located in the promoter regions near the transcription start sites (TSS) of over 60% of human genes. Unlike the majority of CpG sites throughout the genome, the sites located within CGIs tend to be protected from methylation in healthy cells. Therefore the genome-wide DNA methylation pattern in all cells of the body is basically bimodal, with the large majority of CpG sites modified at high levels and CGIs largely unmethylated.

Histone code: an overview

Chromatin is made of repeating units of nucleosomes, which consist of ~146 base pairs of DNA wrapped around an octamer of four core histone proteins (H2A, H2B, H3, and H4) [6]. Histone proteins contain a globular C-terminal domain and an unstructured N-terminal tail [6]. The N-terminal tails of histones can undergo a variety of post-translational covalent modifications including methylation, acetylation, ubiquitylation, sumoylation and phosphorylation on specific residues [7]. The complement of modifications is proposed to store the epigenetic memory inside a cell in the form of a 'histone code' that determines the structure and activity of different chromatin regions [8]. Histone modifications work by either changing the accessibility of chromatin or by recruiting and/or occluding non-histone effector proteins, which decode the message encoded by the modification patterns. The mechanism of inheritance of this histone code, however, is still not fully understood. Unlike DNA methylation, histone modifications can lead to either activation or repression depending upon which residues are modified and the type of modifications present. For example, lysine acetylation correlates with transcriptional activation [9], whereas lysine methylation leads to transcriptional activation or repression depending upon which residue is modified and the degree of methylation. For example, trimethylation of lysine 4 on histone H3 (H3K4me3) is enriched at transcriptionally active gene promoters [10], whereas trimethylation of H3K9

(H3K9me3) and H3K27 (H3K27me3) is present at gene promoters that are transcriptionally repressed [7]. The latter two modifications together constitute the two main silencing mechanisms in mammalian cells, H3K9me3 working in concert with DNA methylation and H3K27me3 largely working exclusive of DNA methylation. Specific patterns of histone modifications are present within distinct cell types and are proposed to play a key role in determining cellular identity [11]. For example, embryonic stem (ES) cells possess 'bivalent domains' that contain coexisting active (H3K4me3) and repressive (H3K27me3) marks at promoters of developmentally important genes [12]. Such bivalent domains are established by the activity of two critical regulators of development in mammals: the polycomb group that catalyzes the repressive H3K27 trimethylation mark and is essential for maintaining ES cell pluripotency through silencing cell fate-specific genes and potentially the trithorax group that catalyzes the activating H3K4 trimethylation mark and is required for maintaining active chromatin states during development [13]. This bivalency is hypothesized to add phenotypic plasticity, enabling ES cells to tightly regulate gene expression during different developmental processes. Differentiated cells lose this bivalency and acquire a more rigid chromatin structure, which may be important for maintaining cell fate during cellular expansion [11]. Histone modification patterns are dynamically regulated by enzymes that add and remove covalent modifications to histone proteins. Histone acetyltransferases (HATs) and histone methyltransferases (HMTs) add acetyl and methyl groups, respectively, whereas HDACs and histone demethylases (HDMs) remove acetyl and methyl groups, respectively [14,15]. These histone-modifying enzymes interact with each other as well as other DNA regulatory mechanisms to tightly link chromatin state and transcription.

In addition, to performing their individual roles, histone modifications and DNA methylation interact with each other at multiple levels to determine gene expression status, chromatin organization and cellular identity [16]. Several HMTs can direct DNA methylation to specific genomic targets by directly recruiting DNA methyltransferases (DNMTs) to stably silence genes [17,18,19]. DNMTs can in turn recruit HDACs and methyl-binding proteins to achieve gene silencing and chromatin condensation [20,21]. DNA methylation can also direct H3K9 methylation through effector proteins, such as methyl-CpG-binding protein 2 (MeCP2), thereby establishing a repressive chromatin state [22].

DNA methylation dynamics in normal cells

Unlike the DNA sequence itself, the bimodal genome-wide DNA methylation pattern is not inherited from the gametes. Rather, it appears that almost all methylation is erased in the very early embryo and a new bimodal pattern is then reestablished at the time of implantation. This 'clearing of the slate' is a key component of the entire epigenetic marking system, as it symbolizes erasure of germ-line programming as a prelude to resetting totipotency. The removal of methyl groups initially begins in the zygote, where specific sequences in the paternal nucleus are actively demethylated [23], and this is followed by more widespread demethylation, which may take place through a combination of active DNA-repair processes together with passive loss of methylation through replication [24]. The function of demethylation before implantation may be important in resetting the genome after gametogenesis and for regenerating totipotency in the preimplantation embryo. However some repetitive sequences as well as other DNA sequences (imprinted regions) retain some of their methylation. The mammalian genome is complex consisting of not only coding sequences but also of transposons and other parasitic elements that have been acquired in the human genome over time. These repetitive sequences make up much of the intergenic and intronic regions of DNA. Many of these repetitive elements contain long terminal repeat promoters which permit the transcription of these sequences [25, 26]. Since the expression of these sequences can allow for the movement of the parasitic elements within the genome, these elements must be persistently silenced by DNA methylation in order to preserve the integrity of the genome [27]. In addition to silencing repetitive elements, CpG methylation in also an important constituent in the establishment and maintenance of imprinted genes in the preimplantation embryo. Gene imprinting is a form of non-Medelian inheritance in which one allele becomes methylated but not in the other leading to mono-allelic expression. Imprinting is important for determining which parental allele will be expressed. Imprinted genes are marked in the gonads by DNA methylation of the imprinted control region (ICR) allowing for the daughter cells to retain the same mono-allelic expression as their parental origin [28]. After erasure of DNA methylation in the early embryo, a new pattern is then established in each individual at about the stage of implantation. This is largely accomplished by the upregulation of the de novo methylases, DNMT3A and DNMT3B, together with DNMT1, which bring about global methylation [29]. Although this active process of de novo methylation appears to be restricted to a short window of time in early development, the resulting pattern of methylation is then maintained

during all subsequent cell divisions [30, 31]. Thus, the bimodal pattern of methylation seen in all somatic cells is a direct reflection of events that occurred at the time of implantation. In coordination with this process, there is also a mechanism for protecting specific sequences, mostly CpG islands [32]. The precise mechanism of CpG island protection is not understood yet. Recently epigenome analyses show that a very high percentage of unmethylated islands contain known transcription start sites, and many specific motifs are also associated with these unmethylated regions, including transcription factor binding sites [30]. Transcription start sites are always packaged in nucleosomes containing H3K4me3 [33,34], and this mark may serve to inhibit the binding of de novo methylases [35,36]. Taken together, these analyses suggest a model whereby the binding of RNA polymerase or other proteins in preimplantation cells may be important in preventing local de novo methylation during the transition to implantation. This implies that the resulting basal methylation pattern simply reflects the potential transcription state of early embryos and, in this way, provides a mechanism to perpetuate this profile in a more stable manner. Although implantation embryos have the capacity to set up the bimodal methylation pattern, the molecular machinery for carrying this out must be downregulated very early in development, as somatic cells are no longer capable of global de novo methylation [37]. Nor do they seem to recognize CpG islands as sites to be protected. Nonetheless, the overall initial pattern formed at the time of implantation is then maintained after each cell division. This is accomplished through the action of DNMT1, which is constantly associated with the DNA replication machinery [38]. This enzyme is highly specific for hemimethylated CpG sites, like those generated during DNA synthesis, and it is this activity that perpetuates the methylation pattern present on the original DNA strands [39]. At the same time, this enzyme has only very low de novo activity, so unmethylated sites remain in this same state during replication. The maintenance of DNA methylation patterns serves an important function during development and aging. In general, gene expression patterns in any cell are determined by two key parameters: the availability of general and specific transcription factors as well as chromatin structure, which modulates local accessibility. Every time cells copy their genetic material as part of the cell division cycle, the replication machinery ‘plows’ through the DNA, thereby disrupting both chromatin structure and factor binding, and these must be rebuilt in every cell generation. In contrast, the underlying DNA methylation pattern is preserved throughout replication, and this serves as a template for guiding the repackaging of DNA without the need to completely rebuild these structures from scratch. In this way, DNA methylation serves as a mechanism for stabilizing gene expression patterns over the entire lifetime of the organism.

After implantation, there are no additional global changes in DNA methylation, and all alterations, whether they involve de novo methylation or demethylation, appear to occur through sequence-specific targeting. One of the first developmental events of this nature is the methylation and silencing of genes responsible for pluripotency, such as Oct3/4 or Nanog [40,41]. Studies in ES cells and in mice have demonstrated that these genes become inactivated in a three-step manner. First, transcription is turned off by direct interactions with repression factors. In the second stage, the histone methylase G9a is recruited to these gene loci; this complex, in coordination with histone deacetylases and H3K4 demethylases, systematically removes all histone activating modifications from local nucleosomes and then brings about methylation of histone H3K9, which in turn binds HP1, leading to formation of heterochromatin. Finally, G9a itself can recruit DNMT3 molecules and cause de novo DNA methylation, an event that occurs with slower kinetics, even in vivo [42,43]. It is clear from this example that DNA methylation itself does not initiate the silencing of pluripotent genes but is rather a secondary or even tertiary effector. This raises the question of what might be the function of de novo methylation in this case. Experiment in ES cells suggest that although DNA methylation is not required for initiating gene silencing, it may be important in maintaining the repressed state over many cell generations, even covering the entire lifespan of the organism. This hypothesis is further supported from a well known example of de novo methylation after implantation: X-chromosome inactivation in female embryos. Random X-chromosome inactivation occurs in each cell concomitantly with differentiation. This process, which is directed by Xist expression on the chosen allele, involves chromosome-wide changes, including a shift to late replication, deacetylation of histones [44], methylation of H3K27 by the Polycomb complex [45,46], and inactivation of many genes. There is no question that the inactivation of X chromosome–linked genes can be accomplished in the absence of DNA methylation as actually occurs in extra-embryonic tissues and in marsupials. The added layer of methylation cells apparently provides long-term stability, making it almost impossible to reactivate genes on the inactive X chromosome in somatic tissues [47]. In contrast, X chromosome–linked genes in marsupials readily undergo derepression over the lifetime of these animals [48]. One of the most interesting concepts that emerged from the targeted de novo methylation is that it is almost always mediated by histone methylases [15] that are recruited by local regulatory factors. As noted above, this is true for *Oct3/4* methylation, which is directed by the H3K9 methylase G9a. Although most CGIs on autosomal genes remain unmethylated in somatic cells, a small number of them (<10%) become methylated in normal tissues and cells [49]. Chromatin immunoprecipitation–high throughput sequencing (ChIP-seq) analysis has shown that a very

high percentage of these CpG islands are actually binding sites for the Polycomb complex, which includes EZH2, a histone methylase specific for H3K27 that can recruit DNMT3 [50]. It appears that these histone-modifying proteins have evolved as a self-contained 'machine' programmed to bring about the epigenetic closure of local gene sequences, and this is accomplished at two different levels: first, by bringing about heterochromatin formation, and then by covalently attaching methyl groups to the DNA, allowing this mark to be stably maintained over many cell generations [15].

An interesting hypothesis that come from the study of de novo methylation predicts that the higher the level of expression is, the less likely it is that a CGI is to become de novo methylated. Direct evidence in support of this prediction has recently come from several exciting papers that have shown that monoallelic methylation of CGIs preferentially occurs on the allele that is less highly expressed. For example, Hitchins et al. [51] showed that an allele of the *MLH1* gene containing a single-nucleotide variant in the promoter, which was less active than the more common allele in transfection experiments, was more likely to become methylated in the somatic cells of cancer-affected families. In other words, the less active allele was the one that was more likely to acquire de novo methylation. An alternative scenario was shown by Boumber et al. [52], who found that an allele of *RIL* (also known as *PDLIM4*) bearing a polymorphism in the promoter that created an additional binding site for the transcription factor SP1 or SP3 was much less likely to become de novo methylated than the allele without this polymorphism. The extra SP1 site therefore confers resistance of this allele to de novo methylation, although the authors could not demonstrate that the extra transcription factor binding site increased gene expression.

Mechanism of gene-specific demethylation

Early studies demonstrated that demethylation is an active event and does not occur passively as a result of DNA replication in the absence of methyl maintenance. For many decades, the biochemical mechanism of demethylation was unknown. Over the past few years, much of the enzymology of demethylation has become clarified. A major advance in our understanding of this process came with the discovery of the Tet family enzymes that can convert 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) [53, 54] and this appears to represent a major intermediate in the demethylation pathway. Demethylation of tissue-specific genes is always associated with activation of these genes during development, but it is not always clear whether demethylation itself is actually required for activation. The

specificity of this process must be directed by factors that recognize nearby cis acting elements, implying that demethylation itself is not the primary event. Rather, it could provide a secondary mechanism for making sure the target gene stably remains in an open conformation. When tissue-specific genes are inserted into a non-expressing cell type by DNA-mediated gene transfer, unmethylated copies are transcribed at a basal level, whereas methylated templates are further inhibited and are expressed at very low levels, similar to that of the parallel endogenous gene in these cells [55, 56]. The same is true for tissue-specific genes that have been programmed to be constitutively unmethylated in transgenic mice [57, 58]. These experiments clearly show that even after bypassing the activation step, one can still observe the effect of undermethylation on long-term expression patterns. Taken together, these studies show that both specific de novo methylation, as well as specific demethylation, operate through similar overall strategies, with targeting being accomplished by interactions between cis-acting sequences and trans-acting recognition factors. In the case of de novo modification, it was demonstrated that methylation enzymes are almost always recruited by local histone methylases [16]. In a similar manner, it is possible that site-specific demethylation of DNA is associated with the presence of histone acetylases or demethylases [59].

Functions of DNA methylation in different genomic contexts: CGIs, start sites and gene bodies

Thanks to improved genome-scale mapping of methylation, we can evaluate DNA methylation in different genomic contexts: transcriptional start sites with or without CpG islands, in gene bodies, at regulatory elements and at repeat sequences. The emerging picture is that the function of DNA methylation seems to vary with context, and the relationship between DNA methylation and transcription is more nuanced than we realized at first. Understanding the functions of DNA methylation requires consideration of the distribution of methylation across the genome.

Patterns at CpG island transcription start sites

Until recently, much of the work on DNA methylation focused on CGIs at transcriptional start sites (TSSs), and it is this work that has tended to shape general perceptions about the function of DNA methylation. The promoter of ~ 60% of human genes falls near a CpG island [60] and the majority are protected from methylation during development and in differentiated tissues. When genes with CGIs at their TSS are actively transcribed, their promoters are usually characterized by nucleosome-depleted regions (NDRs) at the TSS, and these NDRs are often flanked by nucleosomes containing the histone variant H2A.Z and are marked with H3K4me3 [61]. CGI promoters can be repressed by histone modifications and chromatin structure, such as repression mediated by Polycomb proteins. For example, genes encoding master regulators of embryonic development, such as *MYOD1* or *PAX6*, are suppressed by the Polycomb complex both in ESCs and in differentiated cells that are not expressing these genes; they have nucleosomes at the TSS and are marked by H3K27me3, which is generally associated with inactive genes [62]. However, these modifications are easily reversible making them make poor gatekeepers for long-term silencing [63, 64]. Therefore, mammalian cells must possess an additional mechanism for prolong silencing of these sequences. An important component of this process is DNA methylation. When found within promoters, DNA methylation prevents the reactivation of silent genes, even when the repressive histone marks are reversed. Numerous processes by which DNA methylation can influence transcription have been proposed. DNA methylation can directly impede the binding of transcriptional factors to their target sites, thus prohibiting transcription. Other proposed mechanisms are based on the idea that methylation of CpG sequences can alter chromatin structure by effecting histone modifications and nucleosome occupancy within the promoter regions of genes. Many transcription factors are targeted to CG-containing sequences and methylation of CpG sites within these sequences have been shown to prevent the binding of these proteins to these sites [65, 66]. In addition to directly inhibiting transcriptional factors from binding, DNA methylation also recruits methyl binding proteins (MBPs) that specifically bind to methylated CpGs. It has been shown that MBPs can bind repressors and histone deacetylases which may lead to an inactive chromatin structure [67]. Methylation of CpG sites within promoters may also promote nucleosome occupancy at the transcriptional start sites of genes. This in turn could effect transcriptional activation of these genes. Nucleosome occupancy has been shown to decrease the binding of transcription factors and RNA polymerase II [68].

Gene body methylation

Most gene bodies are CpG-poor, are extensively methylated and contain multiple repetitive and transposable elements. Methylation of the CpG sites in gene exons is a major cause of C/T transition mutations, leading to disease-causing mutations in the germline and cancer-causing mutations in somatic cells. It is important to realize that although many CGIs are located at gene promoters, CGIs also exist within the bodies of genes and within gene deserts. Although their functions here remain unknown, Adrian Bird has proposed that these regions may represent 'orphan promoters' that might be used at early stages of development and have escaped methylation in the germline so that their high CpG density is maintained [69]. It has been known from the early days of DNA methylation research that gene body methylation is a feature of transcribed genes [70]. Most gene bodies are not CGIs, and when CGIs are situated in intragenic regions, they were, with a few exceptions [71], thought to remain unmethylated. However, recent experiments have changed this perception: for example, as many as 34% of all intragenic CGIs are methylated in the human brain [72]. The role of this methylation, which is tissue-specific, is not yet clear. It is intriguing, especially because TSSs largely remain unmethylated. Intragenic CGIs can also be preferential sites for de novo methylation in cancer [73]. Even though gene body CGIs can become extensively methylated, this does not block transcription elongation. This is despite the fact that the methylated CGIs are marked by H3K9me3 and are bound by MeCP2, which are chromatin features that are associated with repressed transcription when they are present at the TSS [74]. This leads to an apparent paradox in which methylation in the promoter is inversely correlated with the expression, whereas methylation in the gene body is positively correlated with expression [75]. Thus, in mammals, it is the initiation of transcription but not transcription elongation that seems to be sensitive to DNA methylation silencing. Initially, it was thought that this methylation was primarily a mechanism for silencing repetitive DNA elements, such as retroviruses, LINE1 elements, Alu elements and others, and evidence has been obtained to substantiate this idea [26]. Methylation blocks initiation of transcription at these elements while at the same time allowing transcription of the host gene to run through them. However, whole-genome studies have shown that there might be alternative functions for DNA methylation in gene bodies. This work has shown that exons are more highly methylated than introns, and transitions in the degree of methylation occur at exon–intron boundaries, possibly suggesting a role for methylation in regulating splicing [31].

It is often assumed that TSSs and gene bodies are two separate genomic features. However, most genes have at least two TSSs, so the downstream start sites are within the 'bodies' of the transcriptional units of the upstream promoters. These alternative promoters can be CGIs or non-CGIs, or there can be combinations of an upstream non-CGI and a downstream CGI, or vice versa. These alternative start sites complicate the interpretation of experiments linking expression to methylation, because probes that are used to measure expression often detect the output of all of the promoters, yet only one might be active in a given cell type. Methylation of a downstream promoter would only block transcription from that promoter — it would allow the elongation of a transcript that emanates from an upstream promoter [74]— leading to an apparent discordance between methylation and expression. Indeed, DNA methylation may well be a mechanism for controlling alternative promoter usage [72].

Methylation at enhancers

Enhancers are situated at variable distances from promoters and are key to controlling gene expression in development and cell function. They are mostly CpG-poor, and their methylation status has been examined by whole-methylome analysis. In general, these regions tend to have fairly variable methylation. Stadler et al. [76] identified enhancers in the mouse genome on the basis that they are regions that are not 100% methylated or unmethylated and termed these 'low-methylated regions' (LMRs). Because a given cytosine can either be completely methylated or unmethylated, 'variable methylation' is the outcome of averaging these binary states. This might suggest that the CpG sites are in a dynamic state and that at a given time some are methylated and others are not, owing to competing methylation and demethylation events. Alternatively, the DNA methylation status of each CpG might not be accurately maintained during cell division, and so the LMR state might be due to inefficient inheritance.

Methylation at insulators

Insulators (also known as boundary elements) function to block genes from being affected by the transcriptional activity of neighboring genes. They thus limit the action of transcriptional regulatory elements to defined domains, and partition the genome into discrete realms of expression. Insulators have two main properties: (a) they can block

enhancer-promoter communication (i.e., enhancer-blocking activity), and (b) they can prevent the spread of repressive chromatin (i.e., heterochromatin-barrier activity). For at least some insulators, these two activities can be separable [77]. Typically, insulators are ~ 0.5–3 kb in length, and function in a position-dependent, orientation independent manner. The most well-studied examples are DNA sequences bound by the CTCF protein, which binds to a somewhat heterogeneous sequence motif. A well-studied case is CFCF binding to a site within the imprinted IGF2–H19 locus, at which the presence or absence of CTCF binding controls enhancer–promoter interactions. It has been shown that methylation of a CTCF-binding site at this locus blocks the binding of CTCF, so DNA methylation has an important role in controlling this locus [78].

Aberrant reprogramming of the epigenome in cancer

Tumorigenesis is a multistep process, including initiation, promotion and progression, and a multifactorial pathology characterized by the accumulation of a multitude of alterations including genetic, cytogenetic, and epigenetic changes [79]. Feinberg and Vogelstein reported the first mutation known to result in a human transforming gene: the *c-Ha-ras* oncogene [80]. Since then, a large number of studies have focused on identifying new nonsense, silent DNA and point mutations, deletions, translocations and insertions, and polymorphisms associated with tumor cell growth [81]. High-throughput techniques have been essential for enabling the compilation of a catalogue of rare and common genetic variants. However, a crucial and complementary player in gene regulation — epigenetics — has come to be associated with cancer initiation and development, especially since the recent advent of whole-genome approaches, known as epigenomics. Epigenetic abnormalities in cancer comprise a multitude of aberrations occurring in almost every component of chromatin involved in packaging the human genome [82].

Aberrant DNA methylation was the first epigenetic mark to be associated with cancer as a consequence of the alteration it causes in normal gene regulation [83]. A cancer epigenome is marked by genome-wide hypomethylation, site-specific CpG island promoter hypermethylation and loss of imprinting [84]. While the underlying mechanisms that initiate these global changes are still under investigation, recent studies indicate that some changes occur very early in cancer development and may contribute to cancer initiation. Global DNA hypomethylation plays a significant role in tumorigenesis and occurs at various genomic sequences including repetitive elements, retrotransposons, CpG poor promoters, introns and

gene deserts [85]. DNA hypomethylation at repeat sequences leads to increased genomic instability by promoting chromosomal rearrangements [86]. Hypomethylation of retrotransposons can result in their activation and translocation to other genomic regions, thus increasing genomic instability [87].

In contrast to hypomethylation, which increases genomic instability and activates proto-oncogenes, site-specific hypermethylation contributes to tumorigenesis by silencing tumor suppressor genes. Since the initial discovery of CpG island hypermethylation of the Rb promoter (a tumor suppressor gene associated with retinoblastoma) [88], various other tumor suppressor genes, including *p16*, *MLH1* and *BRCA1*, have also been shown to undergo tumor-specific silencing by hypermethylation [89]. These genes are involved in cellular processes, which are integral to cancer development and progression, including DNA repair, cell cycle, cell adhesion, apoptosis and angiogenesis. Epigenetic silencing of such tumor suppressor genes can also lead to tumor initiation by serving as the second hit required for cancer initiation according to the 'two-hit' model proposed by Alfred Knudson [90]. In addition to direct inactivation of tumor suppressor genes, DNA hypermethylation can also indirectly silence additional classes of genes by silencing transcription factors and DNA repair genes. Promoter hypermethylation-induced silencing of transcription factors, such as *RUNX3* in esophageal cancer [91] and *GATA-4* and *GATA-5* in colorectal and gastric cancers [92], leads to inactivation of their downstream targets. Silencing of DNA repair genes (e.g. *MLH1*, *BRCA1* etc.) enables cells to accumulate further genetic lesions leading to the rapid progression of cancer. While the ability of DNA hypermethylation to silence tumor suppressor genes in cancer is well established, how genes are targeted for this aberrant DNA methylation is still unclear. One possibility is that silencing specific genes by hypermethylation provides a growth advantage to cells resulting in their clonal selection and proliferation. Although it is unlikely that aberrant DNA methylation is strictly a random event and that the methylated genes observed in cancer merely reflect clonal growth advantage resulting from the abnormal methylation, it is likely that many of the methylated genes observed in cancer have been clonally selected. Consistent with the concept of "selective advantage", cancers have tumor specific methylation patterns and the frequency of specific methylated gene varies widely between tumors [93].

Another possibility is that tumor-specific CpG island methylation can occur through a sequence-specific instructive mechanism by which DNMTs are targeted to specific genes by their association with hystone marks. As previously mentioned, DNA methylation and

histone modifications work independently and in concert to alter gene expression during tumorigenesis. A key facet of such silencing mechanisms is the formation of a rigid repressive chromatin state that results in reduced cellular plasticity. The recent discovery of tumor-specific de novo methylation of polycomb target genes, which are silenced by H3K27me3 in normal cells, is another example of this phenomenon [94, 95, 96]. In ES cells, developmentally important genes are reversibly silenced by polycomb proteins through the establishment of the repressive H3K27me3 mark. After differentiation, these genes continue to be repressed through the maintenance of the polycomb mark on their unmethylated promoters by EZH2. In cancer, the polycomb mark is replaced by de novo DNA methylation possibly through the recruitment of DNMTs via the polycomb complex [50]. This tumor specific 'epigenetic switching' of the plastic polycomb mark with more stable DNA methylation results in the permanent silencing of key regulatory genes that may contribute to cell proliferation and tumorigenesis [97]. However, which transformation associated factors trigger this switch is still unclear. The selective hypermethylation of polycomb target genes in cancer cells provide a link between stem cell biology and cancer initiation and a supporting evidence for the 'cancer stem cell' hypothesis. This model suggests that the epigenetic changes, which occur in normal stem or progenitor cells, are the earliest events in cancer initiation. These epigenetic alterations observed in cancers are reflective of the stem cells from which the cells are derived, and that the alterations observed in the cancers merely indicate the undifferentiated state of the tumor cells [98]. The idea that these initial events occur in stem cell populations is supported by the common finding that epigenetic aberrations are some of the earliest events that occur in various types of cancer and also by the discovery that normal tissues have altered progenitor cells in cancer patients. This stem cell-based cancer initiation model is consistent with the observation that tumors contain a heterogenous population of cells with diverse tumorigenic properties [99]. Since epigenetic mechanisms are central to maintenance of stem cell identity [100], it is reasonable to speculate that their disruption may give rise to a high-risk aberrant progenitor cell population that can undergo transformation upon gain of subsequent genetic gatekeeper mutations. Such epigenetic disruptions can lead to an overall increase in number of progenitor cells along with an increase in their ability to maintain their stem cell state, forming a high-risk substrate population that can readily become neoplastic on gain of additional genetic mutations. DNA methylation-induced silencing of genes involved in the regulation of stem/precursor cells' self renewal capacity, such as *p16*, *APC*, *SFRPs* etc., is commonly observed in the early stages of colon and other cancers. Aberrant silencing of these so called 'epigenetic gatekeeper' genes in conditions of chronic stress, such as inflammation, enables stem/

precursor cells to gain infinite renewal capacity thereby becoming immortal. These pre-invasive immortal stem cells are selected for and then form a pool of abnormal precursor cells that can undergo further genetic mutations leading to tumorigenesis [101].

Another common event in carcinogenesis is loss of imprinting defined as the loss of parental allele specific monoallelic expression of genes due to aberrant hypomethylation profiles at one of the two parental alleles. For example, loss of imprinting of *IGF2* has been associated with an increased risk of cancer, including CRC. This event has been observed in different types of neoplasia [102].

Thus, there are many proposed mechanisms by which epigenetic gene regulation is thought to be dysregulated in cancer, and each mechanism is supported by evidence from a variety of different lines of experimental evidence. It is likely that no single mechanism is sufficient to alter the cellular process and drive the pathogenesis of aberrant DNA methylation in cancer. It is most likely that all of these mechanisms contribute to the aberrant epigenetic regulation seen in human cancers, depending on the circumstances of the tumor.

Molecular pathogenesis of sporadic CRCs

CRC is the third most common cancer and the fourth leading cause of cancer-related death worldwide [103]. In 2012, 1.360.600 new cases were diagnosed and 693.900 deaths were attributed to CRC [104]. CRCs occur sporadically in the majority of cases, and only 5%–10% are due to inherited mutations in well-known cancer-related genes. However, up to 25% of patients have a family history of CRC, suggesting a specific contribution by genes that have yet to be identified [105].

Three major genetic pathways for sporadic CRCs

Sporadic CRCs is an heterogeneous disease and evolves through a stepwise accumulation of genetic and epigenetic alterations, leading to the transformation of normal colonic mucosa into invasive cancer. The CRC's heterogeneity reflects the fact that there are many possible etiological pathways responsible for driving CRC development, each of which may be marked by distinct driver mutations and genetic or epigenetic signatures. Importantly, this heterogeneity can also have implications for CRC prognosis and the clinical management of this disease. The conventional model of CRC formation as initially suggested by Fearon and

Vogelstein [106] proposed the adenoma-carcinoma sequence theory, in which APC mutation serves as an initiating event, followed by the accumulation of multiple mutations of genes, such as *KRAS*, *SMAD4*, and *TP53*. According to this model, at least seven distinct mutations are required for CRC pathogenesis.

Presently, three major distinct genetic pathways to CRC have been postulated. Approximately 70% of sporadic CRCs develop through the chromosomal instability (CIN) pathway. These cancers are characterized by the accumulation of numerical or structural chromosomal abnormalities, resulting in aneuploid karyotype, frequent loss-of-heterozygosity (LOH) at tumor suppressor gene loci, and chromosomal rearrangements [107]. Moreover, CIN tumors are distinguished by the accumulation of mutations in specific oncogenes and tumor suppressor genes [e.g., *APC*, *KRAS*, phosphatidylinositol-4,5-bisphosphate 3-kinase, *PIK3CA*, *BRAF*, *SMAD4*, and *TP53*], thereby activating pathways critical for carcinogenesis.

Another important pathway is the microsatellite instability (MSI) pathway, caused by dysfunction of DNA mismatch repair (MMR) genes. MSI is found in 15% of sporadic CRCs. Unlike Lynch syndrome that is caused by germ-line mutations of MMR genes, such as *MLH1* (32% of cases), *MSH2* (39%), *PMS2* (15%), and *MSH6* (14%), MMR deficiency in sporadic CRCs is due mainly to silencing of the MMR genes, mostly *MLH1* (>80% of cases), by promoter hypermethylation [108, 109]. Classification of MSI is based on altered size of various mono and dinucleotide repeat sequences, such as *BAT25*, *BAT26*, *D2S123*, *D5S346*, and *D17S250*, known as the Bethesda panel [110]. Altered size of at least two of the five microsatellite panel markers is defined as MSI-high (MSI-H). Sporadic MSI-H is associated with a third pathway implicated in CRC development namely CpG island methylation phenotype (CIMP). CRCs with one abnormal marker in the panel are termed MSI-low (MSI-L), and their clinical significance is controversial. MSI-L is often grouped with microsatellite-stable (MSS) tumors. Loss or abnormal expression of the MMR proteins MLH1, MSH2, MSH6 and PMS2, assessed by immunohistochemistry, is standard practice in many pathology laboratories and is used to help identify Lynch syndrome along with MSI typing of tumour DNA [111]. Distinguishing Lynch syndrome that show loss of MLH1 expression from sporadic MMR-deficient cancers is currently most appropriately performed by detection of the specific mutation BRAF V600E, which is found in around 80–90 % of sporadic MSI-H CRC, but rarely—if ever—in CRC due to Lynch syndrome [112]. The presence of MLH1 promoter hypermethylation may be used to distinguish sporadic CRC

from Lynch syndrome-associated CRC, but there are interpretative problems as constitutive MLH1 promoter methylation may occur, as well as technical challenges of performing this test [113].

The third pathway, designated as CpG island methylation phenotype (CIMP), is characterized by a widespread CpG island methylation [114]. Approximately 30%–40% of sporadic proximal CRCs are CIMP-positive, compared with 3%–12% of distal CRCs [115, 116, 117]. CIMP-positive CRCs often have MSI-H due to methylation of the MLH1 promoter, but more than 50% of CIMP tumors are MSS. CIMP is uncommon in Lynch syndrome that exhibits MSI [118]. CIMP is also associated with BRAF mutations in both MSI and MSS CRCs [119]. No consensus exists yet for what constitutes the optimal panel of CpG sites for CIMP determination. The classic panel consists of CpG sites in *MLH1*, *CDKN2A MINT1*, *MINT2*, and *MINT31* [120]. CIMP positive tumors based on the classic panel can be divided in two types, namely CIMP-high, related to *BRAF* mutations and *MLH1* methylation, and CIMP-low, related to *KRAS* mutations and MSS [121]. CIMP-negative tumors are MSS with frequent *TP53* mutation [122]. Based on a systematic screen of 195 CpG sites, *CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3* and *SOCS1* was proposed as an alternative to the classic panel [109]. CIMP, which was defined by this panel, did not show a relationship to *KRAS*, but did strongly associate with BRAF V600E mutation [123]. The CIMP concept has not been accepted by all researchers in this field, and over the past few years there has been much debate as to whether the CIMP tumours represent a biologically distinct group of CRCs or are an artificially selected group from a continuum of tumours showing different degrees of methylation at particular loci [124].

The definition of the three genetic pathways is not mutually exclusive, as in the case of CIMP, which often results in *MLH1* promoter methylation and MSI. Up to 25% of MSI CRCs can exhibit CIN [125], In addition, whereas CIMP can account for most of the MSI-positive/CIN-negative CRCs, up to 33% of CIMP-positive tumors can exhibit a high degree of chromosomal aberrations [126]. Conversely, as many as 12% of CIN-positive tumors exhibit high levels of MSI [121]. The significance and implications of these overlapping features are not yet fully defined.

Consistent with genetic models, there appear to be at least three distinct clinicopathologic evolutional routes to sporadic CRCs [127, 128]. The first is the traditional pathway, which starts from normal mucosa via tubular adenomas (with *APC* mutations) and results in typical CRC in the distal colon (with *TP53* mutation and CIN). The second is the serrated pathway,

which starts from normal mucosa via serrated adenomas (with *BRAF* mutations and CIMP) and results in colon cancer in the proximal colon with good prognosis (with *MLH1* loss and MSI). The third is the alternative pathway, which starts from normal mucosa via villous, partly serrated adenomas (with *KRAS*, *BRAF*, and *APC* mutations and CIMP) and results in colon cancer with poor prognosis (with CIMP). The traditional and serrated pathways are homogenous, but the alternative pathway is more heterogeneous. The prevalence of each pathway is estimated at 50%–70% (traditional), 10%–20% (serrated), and 10%–30% (alternative) [128].

Genome-wide detection methodologies for DNA methylation

There are over 28 million CpG sites in the human genome. Assessing the methylation status of each of these sites will be required to understand fully the role of DNA methylation in health and disease. A wide range of experimental methods in a genome-wide scale have been developed to generate quantitative and qualitative information on DNA methylation. Generally, all of the methods include two procedures: the methylation-dependent pretreatment of the DNA and the following analytical step. There are mainly three kinds of pretreatment approaches: enzyme digestion, affinity enrichment and bisulfite conversion.

Whole-genome bisulfite sequencing

Bisulfite-sequencing, which was developed by Frommer and Clark [129], is considered the ‘gold standard’ for DNA methylation analyses. DNA is treated with sodium bisulfite to convert cytosine to uracil, which is converted to thymine after PCR amplification, whereas 5MeC residues are not converted and remain as cytosines [130]. To perform whole-genome bisulfite sequencing (WGBS), genomic DNA (1–5 mg) is sheared and ligated to methylated adaptors before size selection and bisulfite conversion, followed by library construction and highthroughput sequencing. More than 500 million paired-end reads are required to achieve approximately 30-fold coverage of the 28 217 009 CpG sites on autosomes and sex chromosomes; typically approximately 95% of all CpG sites in the genome can be assessed using WBGS. WGBS is fairly accurate and reproduciblehas and has the advantage of providing single nucleotide resolution and whole-genome coverage. However, it typically requires relatively large quantities of DNA (1–5 ug) and accurate interpretation requires

computational expertise. In addition, its high cost makes it unpractical to be applied to large sample size.

Enrichment-based technologies

Genome-wide affinity-based methods rely on enrichment of methylated regions. Two of the common enrichment approaches include methyl-DNA immunoprecipitation (MeDIP), which uses a monoclonal antibody specific for 5-methylcytosine [131] and affinity capture with MBDCap proteins [132, 133]. Both MeDIP and MBDCap can be combined with next-generation sequencing (MeDIP-Seq and MBDCap-Seq). MBDCap-Seq is one of the most widely used capture approaches. The workflow for MBDCap-Seq exhibits similarities to WGBS, but is devoid of a bisulfite conversion step. To perform MBDCap-Seq, genomic DNA (0.2–1 mg) is sonicated before capturing methylated DNA with MBD protein coupled to streptavidin beads. Following capture, the bound methylated DNA can be eluted as a single fraction or in a step-wise elution series to enrich different CpG densities. Enriched DNA is then subjected to library preparation and high-throughput sequencing. Approximately 30 million single-end reads are required for accurate interpretation of data. MBDCap-Seq performed on fully methylated DNA can yield approximately 18% coverage of the genome because it captures approximately 5 million methylated CpG sites. MBDCap-seq is a simple approach that does not require bisulfite conversion and can be used to identify differentially methylated regions [134, 135]. However, a notable disadvantage of MBDCap-Seq is that it does not provide single-nucleotide resolution. Rather, it identifies regions containing multiple methylated CpG sites typically at CpG-rich regions in a readout similar to chromatin immunoprecipitation (ChIP-Seq).

Reduced representative bisulfite sequencing

Reduced representative bisulfite sequencing (RRBS) is an efficient and high-throughput technique used to analyze methylation profiles at a single-nucleotide level from regions of high CpG content (e.g., CpG islands), but does not interrogate intergenic or lowly methylated regions of the genome [136]. RRBS relies first on the digestion of genomic DNA (0.01–0.03 mg) with a methylation-insensitive restriction enzyme, such as MspI, that selects genomic regions with moderate to high CpG density, such as CpG islands, followed by DNA size fractionation. This ‘reduced representation’ of the genome is sequenced similarly to

WGBS to generate a single-base pair resolution DNA methylation map [136]. A minimum of approximately 10 million sequencing reads are required for the downstream analysis of RRBS data sets, leading to approximately 3.7% actual coverage of CpG dinucleotides genome-wide or approximately 1 million CpG sites. One of the main advantages of RRBS is that it is more cost-effective than WGBS, because it targets bisulfite sequencing to an enriched population of the genome, while retaining single-nucleotide resolution. RRBS data are restricted to regions with moderate to high CpG density, and are enriched for promoter-associated CpG islands. However, RRBS interrogates only ~ 4% of the approximately 28 million CpG dinucleotides distributed throughout the human genome. Thus, a lack of coverage at intergenic and distal regulatory elements is a potential disadvantage of the method.

Infinium HumanMethylation450 BeadChip

The Infinium HumanMethylation450 BeadChip (450K) is an attractive option for genome-wide DNA methylation analyses in a variety of cell types. It is suitable for clinical samples, it requires little starting material (approximately 0.5 mg), is cost effective, and can be used in a high-throughput manner. The 450K protocol begins with the bisulfite conversion of genomic DNA (0.5–1 mg). Converted genomic DNA is hybridized to arrays that contain predesigned probes to distinguish chemically methylated (cytosine) and unmethylated (converted to uracil). A single-base extension step incorporates a labeled nucleotide that is fluorescently stained. Scanning of the array detects the ratio of fluorescent signal arising from the unmethylated probe compared with the methylated probe, allowing the level of methylation to be determined. The 450K BeadChip interrogates 482422 cytosines across the human genome, which represents only approximately 1.7% of all CpG sites in the human genome, substantially less than other methods. However, these sites are enriched for CpG (99.3%) residues and almost half (> 41%, approximately 197790 CpG sites) of the probes on the array cover intergenic regions, such as bioinformatically predicted enhancers, DNase I hypersensitive sites, and validated differentially methylated regions (DMRs) [137, 138]. Therefore, 450K has become the method of choice for genome-wide DNA methylation analyses of profile large cohorts, because it requires a low amount of input material and it is cost effective. Of note, up to now, thousands of DNA methylation publicly available data sets (e.g TCGA) [139] have been generated from this array-based detection method and these data have been widely used to infer the candidate biomarkers for cancer diagnosis. However,

when using 450K BeadChip technology, there are also some issues to consider. First, the design is heavily biased due to preselection and inclusion of probes that interrogate only certain CpG sites that have been previously identified in methylation-based assays and, therefore, the design is not hypothesis neutral. Second, it is assumed that CpG sites located adjacent to those interrogated by the probes will be similarly un/methylated, which is known as the 'co-methylation assumption' [140]. Finally, there are behavioral differences between the two types of probe design on the array, and the filtering of probes may be affected by single nucleotide polymorphisms, which need to be factored in to the data analysis pipelines [141].

Processing and analysis methods for Infinium HumanMethylation450 BeadChip

The 450K arrays are based on the Infinium chemistry and represent an extension of the previous Infinium Human Methylation27 BeadChip (27K) platform, which was biased toward promoter regions. This extension resulted in wider coverage, specially toward other genomic regions like gene bodies and CpG shores. However, this also resulted in the introduction of two different bead types associated to two different chemical assays, Infinium I and Infinium II. Infinium I consists of two bead types (Methylated and Unmethylated) for the same CpG locus, both sharing the same color channel, whereas Infinium II utilizes a single bead type and two color channels (green and red) [137]. Infinium II assays have larger variance and are less sensitive for the detection of extreme methylation values, which is probably associated to the dual-channel readout, thus rendering the Infinium I assay a better estimator of the true methylation state [142, 143, 144]. Moreover, different genomic elements (promoters, CpG islands, gene bodies, etc.) have different relative fraction of type I or type II probes [144]. The inclusion of two different bead types (Infinium I and Infinium II) introduce probe-type bias complicating the analysis of the 450K arrays. Different statistical methods has been developed to correct the bias due to the two different chemical assays.

The cytosine methylation status for single CpG sites at each allele is always binary (0 or 1). However, the measured methylation levels can, in principle, take any value between 0 and 1 when averaging over many cells, or when the methylation status differs between the two alleles (imprinting, X-chromosome inactivation). For bisulfite microarrays, the methylation level is usually measured in two different scales, the β value and the M-value.

The β value is calculated as $\beta = M/(M + U + \alpha)$ where M and U are methylated and unmethylated signal intensities and α is an arbitrary offset (usually 100) and can be interpreted as the percentage of methylation (it ranges from 0 to 1). The alternative index is not bounded by 0 and 1 and is calculated as $M = \log2((M + \alpha)/(U + \alpha))$, which is essentially equivalent to a logit transformation of β. Even if M-values cannot be directly interpreted as methylation percentages, they offer several advantages, including the possibility of employing downstream association models that rely on the assumption of Gaussianity, as β values appear compressed in the high and low range and often display heteroscedasticity. However, from a pragmatic point of view and to allow biological interpretation, it is always advisable to report the final effect size in terms of median or mean β value change, even if the feature selection step has been performed in the M-value space.

Normalisation

Normalisation concerns the removal of sources of experimental artifacts, random noise and technical and systematic variation caused by microarray technology, which, if left unaddressed, has the potential to mask true biological differences [145]. Two different types of normalisation exist: (1) within-array normalisation, correcting for intensity-related dye biases, and (2) between-array normalisation, removing technical artifacts between samples on different arrays [146].

Within-array normalisation: probe-type normalization

Independently of the scale used (β value or M-value), the methylation profile for each sample shows a bimodal distribution, with two peaks corresponding to the unmethylated and methylated CpG positions. Because of the technical differences in probe design, a correction method is advisable. Specifically, the 450K array has 485577 probes, of which 72% use the Infinium type II primer extension assay where the unmethylated (red channel) and methylated (green channel) signals are measured by a single bead. The remainder use the Infinium type I primer extension assay (also used in the 27K) where the unmethylated and methylated signals are measured by different beads in the same colour channel. Importantly, the two probes differ in terms of CpG density, with more CpGs mapping to CpG islands for type I probes (57%) as compared with type II probes (21%). Moreover, compared with Infinium I probes, the range of β-values obtained from the Infinium II probes is smaller; in

addition, the Infinium II probes also appear to be less sensitive for the detection of extreme methylation values and display a greater variance between replicates [137, 144]. The divergence in the methylation distribution range has implications for statistical analysis of the array data. For example, in a supervised analysis of all probes, an enrichment bias towards type I probes may be created when ranking probes because of the higher range of type I probes [147]. Attempts have been made to use rescaling to 'repair' the divergence between these two types of probes. Methods for reducing the probe-type bias include a peak-based correction [144], SWAN method [147], subset quantile normalization [148], and BMIQ [143]. In a benchmarking work [142], BMIQ resulted as the best algorithm for reducing probe design bias. BMIQ, which employs a betamixture and quantile dilation intra-array normalization strategy, is available through several R packages, RnBeads [149], WateRmelon [150].

Between-array normalisation

Between-array normalization is intended to remove part of the technical variability that is not associated with any biological factor, but which can be considered as caused by experimental procedures. Specifically, there is an imbalance in methylation levels throughout the genome creating a skewness to the methylation log-ratio distribution [146]. This imbalance is due to the non-random distribution of CpG sites throughout the genome and the link between CpG density and DNA methylation; for instance, CGI are often unmethylated, whereas the opposite relationship is typically seen in non-CGIs in normal human cells [151]. Owing to this features of DNA methylation, there is a lack of consensus regarding the optimal approach for normalisation of methylation data although a comparison of different normalization pipelines has been performed in recent works [142]. Many of the proposed approaches employ a form of quantile normalization (QN), which has been shown to perform well for gene expression studies. The goal of QN is to produce identical distribution of probe intensities for all the arrays and it has been applied to 450K data in several forms [152]. While forcing the distribution of the methylation estimates to be the same for all the samples is a reasonably too strong an assumption for many biological comparisons, normalizing signal intensities appears a valid alternative in reducing technical variability in several contexts [142, 152]. However, examination of the signal intensities and the study design should guide the application of this level of between-samples normalization, in order not to harm the integrity of the biological signal. A recent extension of QN, termed functional

normalization [153], uses control probes from the array to remove unwanted variation, assuming that summarized control probes function as surrogates of the nonbiological variation, which may include batch effects. Several comprehensive R packages have been developed for the processing and the analysis of 450K data such as lumi [154], methylumi [155], minfi [156], wateRmelon [150], ChAMP [157], and RnBeads [149].

Another type of unwanted variation in 450K data is represented by batch effects, which contaminate many high-throughput experiments including 450K arrays. A batch is defined as a subgroup of samples or experiments exhibiting a systematic non-biological difference that is not correlated with the biological variables under study. For example, different batches are represented by groups of samples that are processed separately, on different days or by a different operator. Batch effects can only affect a subset of probes instead of generating artifacts globally; therefore, many normalization methods fail in eliminating or reducing batch effects. Specific methods have been developed to deal with this source of variability, including ComBat [158] and SVA (Surrogate Variable Analysis) [159]. These methods aim at removing the unwanted variation that remains in high-throughput assays despite the application of between-sample normalization procedures. ComBat method rely on the explicit specification of the experimental design, in order to maintain the variability associated to a biological factor, while removing variability associated to either known or unknown batch covariates. The ComBat method directly removes known batch effects and returns adjusted methylation data, by using an empirical Bayes procedure. However, when the sources of unwanted variation are unknown, surrogate variables can be identified by SVA directly from the array data. It is important to remember that the best safeguard against problematic batch effects is a careful experimental design, coupled with a random assignment of the samples to the arrays, the inclusion of a method to account for batch effect and possibly the presence of technical replicates, one for each processing subgroup, if the samples cannot be processed together in the case of large cohorts.

Biomarkers

A biomarker is any biological characteristic that can be objectively measured and evaluated as an indicator of normal biological process, pathogenic process, or pharmacological response to a therapeutic intervention [160]. Biomarkers can be used at any stage of a disease and can be associated with its cause or latency (risk biomarkers), onset (diagnostic biomarkers), clinical course (prognostic biomarkers), or response to treatment

(predictive biomarkers) [161, 162]. Biomarkers can also be associated with specific environments (exposure biomarkers). As almost all complex human diseases are caused by a mixture of genetic and environmental variation, biomarkers, especially those antecedent to disease, can be influenced by either of these factors. Biomarkers can also reflect the mechanisms by which exposure and disease are related. They can stratify individuals according to risk or prognosis and they can be used as targets or surrogate endpoints in clinical trials. An ideal biomarker must be able to provide clinically-relevant information, be accurately measurable in multiple individuals, ideally across multiple populations. Almost any biological tissue sample or bodily fluid can be used for DNA methylation analysis. DNA methylation is the most robust epigenetic mark and will survive most sample storage conditions. The robustness of DNA methylation marks makes DNA methylation analysis very attractive in a clinical environment for the early detection of cancer and easy-to-access tissues or bodily fluids can be collected. Such samples include venous peripheral blood, buccal epithelium or saliva, urine, stools, bronchial aspirates, and, even in some cases, muscle or adipose tissue.

Colon cancer can be cured by a relatively simple surgical procedure when the cancer is diagnosed early before metastasis occurs. CRC is suitable for early detection approaches due to its recognizable early stage and its defined natural history. The progression from an adenoma to carcinoma may take decades, which provides a window of opportunity for early CRC detection. To reduce disease-specific mortality, it is therefore important to identify and treat CRC as early as possible. Mass screening would therefore greatly contribute to the early diagnosis and timely treatment of CRC. Among screening tests for CRC, colonoscopy and fecal occult blood test (FOBT) are used most frequently; the former is highly sensitive but often requires hospitalization of the patient and is, therefore, costly, while the latter is relatively simple to use but has a low positive predictive value. The FOBT is an additional screening method that has been shown to be a highly cost-effective, noninvasive screening method, reducing CRC-related mortality [163]. The FOBT checks for non-visible “occult” blood in the stool of patients. However, although the performance of the immunochemical FOBT has been improved, and is now widely used in Europe for CRC screening, FOBT remains limited in the detection of early-stage CRC [164]. Therefore, there is an urgent need to develop simple and less invasive tests with high sensitivity and specificity.

For the diagnosis of colon cancer, markers that have high sensitivity and specificity are essential and initiating cancer screening programs is the very first step in reducing cancer-

related mortality. Patients are more willing to participate with less invasive screening methods. Epigenetic alterations of specific genes have recently emerged as potential candidate biomarkers for the early detection of cancer [165]. The most easily accessible sources to study DNA methylation are bodily fluids, such as blood, stool or urine. It is therefore interesting to discover novel non-invasive biomarkers with diagnostic utility for the detection of CRC. Currently, many highly methylated markers have been reported in CRC, but only Vimentin (*VIM*) and *SEPT9* are included in commercial non-invasive tests [166].

Markers in stool samples

Detection of tumor-derived DNA alterations in stool is an intriguing new approach with a high potential for the noninvasive detection of CRC. Tumors release markers at different stages of progression, by different mechanisms, into different media that can be assayed. Tumor cells and most tumor markers likely enter into stool at earlier stages than into blood or urine, an advantage of stool testing for cancer precursor lesions and early-stage tumors. Dysplastic cells and their constituents are released into stool by exfoliation from the surface of precancerous lesions and early-stage cancers. Exfoliation from colorectal neoplasms appears to be a continuous process that occurs more frequently than exfoliation from normal epithelium [167]. Factors that might contribute to the high rate of exfoliation from tumors include increased proliferation and reduced cell-cell or basement membrane adhesion. In normal colon, epithelial renewal does not necessarily lead to exfoliation, but instead, often involves engulfment of effete colonocytes by sub-epithelial phagocytes[168]. Recent studies have identified an increasing number of genes that are methylated in stool samples of CRC patients and *VIM* is unique and particularly interesting. The expression of *VIM* does not seem to be under epigenetic regulation since the gene is not methylated and is transcriptionally silent in normal colorectal epithelial crypt cells. *VIM* has been thus rarely considered as a target for cancer-associated aberrant methylation. However, the usefulness and potential of the *VIM* gene as a methylation marker in CRC has recently emerged. Chen et al. [169] found that a CpG island of *VIM*, located upstream of the first exon and normally unmethylated, became densely methylated in CRC and was significantly associated with Dukes' stage with a trend toward preferentially developing liver metastasis and peritoneal dissemination. Aberrant *VIM* methylation can be detected in fecal DNA from CRC patients [170], but rarely in normal colon tissues and control fecal DNA from healthy subjects. Therefore, *VIM* methylation might be useful in identifying individuals with colon cancer [171]. As

mentioned, the detection of aberrantly methylated genes in fecal DNA has potential value in the noninvasive diagnosis of colorectal neoplasms. For this purpose, the novel identification of genes that are frequently methylated in cancer in stool samples from CRC patients requires high detection sensitivity/specificity for clinical use. Detection of patients with colon polyps and larger adenomas in stool samples by testing gene methylation may improve overall sensitivity. Current approaches to test gene methylation in fecal DNA samples need to be further developed because of false-negative and false-positive cases which currently limit the detection accuracy [172].

Markers in serum/plasma

Tumor cells gain entry into blood via blood vessel invasion which occurs in cancers but not precancerous lesions. Histological analyses have shown that blood vessel invasion occurs more frequently from advanced than early-stage tumors [173] and that there is more abundant release of tumor cells into the circulation with advanced cancers[174]. Therefore circulating tumor cells in peripheral blood (PB) may reflect certain biological characteristics of tumors which in turn may predict the potential of tumor metastasis and recurrence[175]. Tumor markers can also enter blood indirectly via inflammatory cells that infiltrate tumors, phagocytose dysplastic cells (part of the immune response), and then re-enter the circulation carrying detectable patterns of tumor-derived nucleic acids or proteins. This alternative route of circulatory marker release via phagocytic leukocytes can occur during all stages of tumorigenesis and could potentially allow for detection of precancerous lesions as well as cancers by a blood test. Conventional cancer markers such as carcinoembryonic antigen (CEA) were developed by quantifying small amounts of circulating proteins. These markers are specific for certain types of cancer, having proven to be of some value in the early detection of cancers and cancer relapse, monitoring the response of cancers to therapy, and as predictors of cancer prognosis [176]. However, the shortcomings of this approach due to the limited sensitivity and specificity are now well-recognized [177]. The measurements of serum CEA levels are now used to monitor disease progression and response to therapy in patients with CRC [178]. It is the only serum marker that is recommended to be added to the established tumor–node– metastasis staging system [179] due to its prognostic significance in Dukes' B or equivalent stages. However, only a proportion of CRC express elevated CEA levels at the time of diagnosis. In the search for increasing the pool of useful serologic markers, the potential use of nucleic acid markers in plasma and serum has been examined.

PCR-based assays in small amounts of nucleic acids can thereby detect and quantify genetic and epigenetic alterations in the circulating tumor DNA [180]. Although it is widely accepted that DNA methylation markers might increase cancer detection at earlier stages and eventually contribute to future advances in assessing clinical outcome of cancer patients, methylation analysis of a number of gene promoters in DNA from blood samples has been limited. Despite great efforts in detected of diagnostic/prognostic biomarkers for CRC, since 1997, there is not any reliable serum biomarker that could be used as a noninvasive screening method. A SEPT9 represent a promising blood-based biomarker assay developed and validated in case-control studies by deVos et al. [181]. The performance of the SEPT9 assay was examined in a study of 97 cases with CRC and 172 healthy controls. The SEPT9 assay yielded a sensitivity of 72% at a specificity of 92% in the training study and 68% sensitivity at a 89% specificity in the testing study. The authors concluded that circulating methylated SEPT9 DNA is a valuable biomarker for minimally invasive detection of CRC and could be implemented in a standardized assay.

Technical aspects in biomarker research and in its clinical application

Research into methods for noninvasive molecular detection of colorectal neoplasia is a continuously changing field. Technological advances that have improved test performance include innovative methods to increase analytical sensitivity, development of buffers that prevent marker degradation during transport and storage, and identification of marker panels that effectively cover the various genotypes of colorectal neoplasms. Development of high throughput platforms should expand the capacity of these assays and lower costs. Tests to reliably detect the minute quantities of marker analytes in stool, blood, and urine must have high levels of sensitivity, especially if precancerous lesions and small, early-stage tumors are to be identified. The technical challenges differ depending on the medium and the markers tested. In stool samples, it is a challenge to detect trace amounts of target DNA among large amounts of background DNA and high analytical sensitivity (ie, reliable detection of low analyte concentrations) is required. Human DNA concentrations average about 100 ng/g, which is roughly 0.01% of the total stool DNA. The other 99.99% of stool DNA is nonhuman, mostly bacterial and some dietary. The mutated or aberrantly methylated copies of the tumor genes to be identified are only a small proportion of the minute fraction of stool DNA that is of human origin[182, 183]. Accordingly, an enrichment step is often needed to capture target gene sequences for use as a polymerase chain reaction (PCR) template and

remove PCR inhibitors before the assay is performed. A number of new approaches have substantially improved analytical sensitivity. For example, Methyl-BEAMING and a digital melt curve method[182, 183, 184] detect <0.1% of mutant copies, providing the requisite analytical sensitivity for detection of precursor lesions. In contrast to stool, essentially all DNA in a plasma or serum sample is of host origin; tumor-derived DNA can account for >25% of total circulating DNA levels[185]. Although the minimal amount of background DNA in plasma or serum may enhance assay discrimination, the amount of altered DNA is often absent or below detectable limits from patients with precancerous lesions or early-stage tumors[185]. As with stool and urine, optimized sample processing and removal of PCR inhibitors are required for high levels of sensitivity[185].

Markers are often degraded during specimen transport or storage, which can reduce test sensitivity. Adding stabilizing buffers to the biospecimen at the time of collection can eliminate or substantially reduce marker degradation. Addition of buffers containing DNAase inhibitors effectively prevents marker degradation during transport and storage[186, 187]. As assays are developed, it will be important to include specifically designed stabilization approaches, based on the markers and medium tested.

Aims

The aims of the present work were:

- To identify signature alterations in CRC methylome
- To test whether these alterations represent early events in CRC development
- To explore the use of non-invasive techniques (stool and ctDNA) to reveal altered methylation
- To correlate the mRNA gene expression of CRCs with the altered DNA methylation

Materials and methods

Samples for whole genome methylation analysis

The methylome analysis was first performed in 18 CRCs among which four had matched peritumoral samples (selected to represent the four anatomic region affected by cancer: left, right, sigmoid colon, rectum). Tissue samples were collected at the Department of Surgical Sciences, University of Cagliari (Italy). Table 1 describes the clinical features of the analysed samples.

Table 1: CRC samples used for the methylome study

	KRAS	MSI	Dukes	Histology	CIMP_like	Anatomic site
254_P	Wt	Wt		Peritumoral	CIMP_neg	Left Colon
264_P	Wt	Wt		Peritumoral	CIMP_neg	Sigmoid colon
279_P	Wt	Wt		Peritumoral	CIMP_neg	Rectum
359_P	Wt	Wt		Peritumoral	CIMP_neg	Right Colon
254_T	Wt	Wt	b	CRC	CIMP_neg	Left Colon
264_T	Wt	Wt	b	CRC	CIMP_neg	Sigmoid colon
279_T	Mutated	Wt	d	CRC	CIMP_neg	Rectum
308_T	Wt	Wt	b	CRC	CIMP_neg	Sigmoid colon
309_T	Wt	Wt	a	CRC	CIMP_neg	Sigmoid colon
310_T	Wt	Wt	a	CRC	CIMP_pos(>20%)	Left Colon
311_T	Wt	Mutated	a	CRC	CIMP_neg	Rectum
313_T	Wt	Wt	b	CRC	CIMP_pos(>20%)	Sigmoid colon
325_T	Wt	Wt	d	CRC	CIMP_neg	Rectum
337_T	Wt	Wt	d	CRC	CIMP_pos(>20%)	Left Colon
352_T	Wt	Wt	d	CRC	CIMP_neg	Sigmoid colon
359_T	Mutated	Wt	b	CRC	CIMP_pos(>30%)	Right Colon
362_T	Wt	Wt	a	CRC	CIMP_neg	Rectum
368_T	Mutated	Wt	b	CRC	CIMP_neg	Sigmoid colon
376_T	Wt	Mutated	b	CRC	CIMP_pos(>30%)	Right Colon
400_T	Mutated	Wt	d	CRC	CIMP_neg	Rectum
407_T	Mutated	Wt	d	CRC	CIMP_neg	Sigmoid colon
455_T	Wt	Wt	d	CRC	CIMP_neg	Sigmoid colon

In a second step, methylome analysis was conducted in 21 adenomas and three matched intestinal mucosa controls, from 21 patients bearing an adenoma. Lesions were removed during endoscopy (De Benedetti et al. 1994). DNA samples were collected at the National Institute for Cancer Research of Genoa (Italy). Table 2 describes the clinical features of the analyses samples.

Table 2: Adenomas samples used for the methylome study

	KRAS	APC	Grade	Histology	CIMP_like	Anatomic site
CTE1279	Wt	Wt		Normal	CIMP_neg	Right Colon
CTE1434	Wt	Wt		Normal	CIMP_neg	Left Colon
CTE1620	Wt	Wt		Normal	CIMP_neg	Left Colon
CTE1266	Wt	Mutated	Adenoma_mild_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1280	Mutated	Wt	Adenoma_mild_dysplasia	Adenoma	CIMP_pos(>30%)	Right Colon
CTE1435	Mutated	Wt	Adenoma_severe_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1470	Wt	Wt	Adenoma_low_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1473	Mutated	Wt	Adenoma_mild_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1474	Mutated	Wt	Adenoma_mild_dysplasia	Adenoma	CIMP_neg	NA
CTE1540	Wt	Wt	Early cancer in adenoma	Adenoma	CIMP_neg	Left Colon
CTE1619	Wt	Wt	Adenoma_mild_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1621	Wt	Mutated	Adenoma_severe_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE1727	Wt	Mutated	Early cancer in adenoma	Adenoma	CIMP_neg	Left Colon
CTE1730	Wt	Wt	Early cancer in adenoma	Adenoma	CIMP_neg	Left Colon
CTE1748	Mutated	Wt	Early cancer in adenoma	Adenoma	CIMP_neg	Left Colon
CTE1877	Wt	Wt	Early cancer in adenoma	Adenoma	CIMP_neg	Left Colon
CTE2032	Wt	Wt	Adenoma_mild_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE2034	Wt	Mutated	Adenoma_low_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE2035	Wt	Wt	Adenoma_low_dysplasia	Adenoma	CIMP_neg	Right Colon
CTE2036	Wt	Wt	Hyperplastic polyp	Adenoma	CIMP_neg	Left Colon
CTE2040	Wt	Wt	Adenoma_low_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE2046	Wt	Wt	Adenoma_severe_dysplasia	Adenoma	CIMP_neg	Right Colon
CTE2052	Wt	Wt	Adenoma_low_dysplasia	Adenoma	CIMP_neg	Left Colon
CTE2055	Wt	Wt	Hyperplastic polyp	Adenoma	CIMP_neg	Left Colon

DNA extraction and bisulphite treatment of the DNA

Genomic DNA was extracted from tumoral and peritumoral tissue using the DNeasy Blood & Tissue Kit (Qiagen). Quality control and quantification of DNA were performed before and after bisulphite conversion. DNA was quantified with NanoDrop (NanoDrop Products Thermo Scientific Wilmington, DE) and by fluorometric reading (Quant-iT™ PicoGreen® dsDNA Assay Kit); quality was assessed by visualization of genomic DNA on 1% agarose gel electrophoresis. Only DNA samples not fragmented and with a concentration higher than 50 ng/μl were subsequently processed. The genomic DNA was treated with sodium bisulfite using the EZ DNA Methylation Kit ™ (Zymo Research); the technique requires only 500 ng of input DNA.

DNA methylation assay

Four microliters of bisulfite-converted DNA were used for hybridization on Infinium HumanMethylation 450 BeadChip, following the Illumina Infinium HD Methylation protocol. Data were acquired on an Illumina HiScan SQ scanner. Image intensities were extracted using GenomeStudio (2010.3). The methylation score for each CpG site was represented as β values according to the fluorescent intensity ratio between methylated and unmethylated probes. Same procedure has undergone the DNA extracted from 21 adenomas and three normal mucosae.

Data management, quality control, preprocessing, normalisation and annotation

Illumina methylation 450K raw data were analysed using the R/Bioncondictor package "RnBeads". RnBeads is an R/Bioconductor package for the comprehensive analysis of genome-wide DNA methylation data with single base-pair resolution. RnBeads builds upon extensive prior research on bioinformatic and statistical methods for DNA methylation analysis and implements a comprehensive analysis pipeline from data import via filtering, normalization and exploratory analyses to characterizing differential methylation. RnBeads is straightforward to run, and the standard pipeline requires an R installation and basic R programming experience. An RnBeads analysis can be launched using a single command in

R: rnb.run.analysis(…), which takes a user-provided sample annotation table as input and extracts relevant information needed to automatically configure the analysis. RnBeads workflows can also be fine-tuned using global configuration parameters, which are specified using rnb.options(...). It is also possible to run some or all steps of the RnBeads workflow interactively and to write R scripts that operate directly on the RnBSet object containing all DNA methylation data and sample annotations of a given analysis.

Illumina methylation 450K data were in the form of Intensity Data (IDAT) files. This is a proprietary format that is output by the scanner and stores summary intensities for each probe on the array. Typically, each IDAT file is approximately 8MB in size. It is recommended to start the analysis from IDAT files to let RnBeads perform quality control for the Infinium 450k data using the microarray control probe information which should be present in the input raw data but not in pre-normalized data (i.e Illumina GenomeStudio files). When IDAT files are loaded into RnBeads, the R/Bioconductor package methylumi is internally used for performing the low-level processing. RnBeads combines the loaded data into a single RnBSet object that constitutes the basis for all further analysis steps. The RnBSet object links the DNA methylation data to genome annotations such as CpG islands, genes and promoters, genome-wide tiling regions and user-defined genomic region sets. The RnBSet object primarily stores DNA methylation levels as beta values, which are used by most modules; nevertheless, RnBeads also calculates M-values and uses them for the limma analysis as part of the differential DNA methylation module. After the DNA methylation data have been loaded in a RnBSet object the next step in the analysis is the quality control (QC) which involves plotting the microarray's quality control probes to monitor different technical parameter such as bisulfite conversion efficiency and unspecific probe hybridization for the detection of technical failures and the evaluation of background signal analysing the distributions of intensities for the approximately 600 negative control probes which are present on the Infinium 450k array. In both channels the negative control probe intensities are expected to be normally distributed around a relatively low mean (as a simple rule, the 0.9 quantile should be below 1000). Any strong deviations from such a picture in one or more samples may indicate quality issues; discarding such samples could be beneficial for downstream analyses. The Infinium 450k BeadChip also contains a small number of genotyping probes that could be used to evaluate the sample mix-ups and to confirm sample identity.

Preprocessing and normalization steps are further required to prepare the methylation data for downstream analysis. Preprocessing the Infinium 450k data, which include probe filtering, it is an important and recommended step to be carried out before and after data normalization to minimize the risk of measurement biases affecting the analysis. In RnBeads this happens in an automated fashion after the user specifies filtering criteria. In our analysis we exclude Infinium probes on sex chromosomes as well as those overlapping with too many SNPs, that stand a high chance of influencing DNA methylation measurements, and sites with too many missing values. Additionally we discarded CpGs that contain a substantial fraction of measurements with low technical quality (e.g., bad detection p-value) using a Greedycut algorithm. The next step is the normalization of Infinium 450k data. RnBeads offers several alternative options for signal intensity-based normalization, which is an important step to reduce probe biases that could interfere with the analysis. We used the SWAN method as implemented in the minfi package which is the RnBeads default for Infinium data normalization. In addition, background subtraction NOOB method as implemented in the methylumi package was applied in combination with the above normalization method.

The case/control differential methylation analysis was conduced using RnBeads default settings. The default differential methylation analysis in RnBeads can be conducted not only at the level of individual CpGs but also by combining measurements across larger genomic regions, which increases statistical power and can result in more interpretable sets of differentially methylated regions. Genomic regions are inferred from the annotation data package, named RnBeads.hg19, that contain annotations for CpG sites, array probes and predefined genomic regions. The default genomic regions taken into account by RnBeads are Ensembl genes (defined as the whole locus from transcription start site to transcription end site), promoters (defined as the regions 1.5 kb upstream and 0.5 kb downstream of transcription start sites), CpG island and tiling regions (5kb windows). In each comparison defined by the sample annotation table, RnBeads initially computes p-values for all covered CpGs. By default, this analysis is performed with hierarchical linear models as implemented in the limma package and using M-values, which exhibit a distribution that is more consistent with limma's statistical model assumptions than the beta values that RnBeads uses in most parts of its analysis. The CpG-level p-values are corrected for multiple testing using the false discovery rate (FDR) method. Furthermore, to obtain aggregate p-values at the level of predefined genomic regions, the uncorrected, CpG-specific p-values within a given region are combined using an extension of Fisher's method. This procedure results in a single

aggregate p-value for each region, and the aggregate p-values are subjected to multiple testing correction using the FDR method.

The results of differential methylation analysis are stored in tabular format, one table for each region considered (CpG sites, genes, promoters, CpG island, tiling), and easily saved in plain text (i.e comma separated values (*.csv)). The subsequent analysis steps are focused on CpG island differential methylation table. While the reasons for this choice are explained in details in the discussion, the following describe how to handle this type of genomic data, from the annotation to gene set/pathways enrichment analysis.

The genome is typically represented as a linear sequence, split over multiple chromosomes, and data are linked to the genome by occupying a range of positions on the sequence. These data fall into two broad categories. First, there are the annotations, such as gene models, transcription factor binding site predictions, GC percentage, polymorphisms, and conservation scores. Second, there are primary experimental measurements, such as percentage of methylation at each CpG locus. Data integration, within and between those two categories, is made possible by treating the data as ranges on the genome, which acts as a common scaffold. Thus, ranges play a central role in genomic data analysis and statistical tools should consider ranges to be as fundamental as quantitative and categorical data types. In R/Bioconductor the packages that form the core of the infrastructure for the integrative statistical analysis of range-based genomic data include IRanges and GenomicRanges. The IRanges package provides the fundamental range data structures and operations but data structures should support the storage of per-range metadata, because genomic data is multivariate and consists of much more than the ranges alone. GenomicRanges builds upon IRanges to add biological semantics, including explicit treatment of chromosome name and strand. The GRanges class supports many of the same range operations as IRanges and specializes them for genomic data. There is a wide set of possible range operations. Certainly, a recurrent operation used is the overlap detection and GRanges method is specifically able to take advantage of the chromosome information when detecting overlaps using the function findOverlaps(…).

Starting from the differential methylation tables of CRC and adenomas and filtering out the most significant results selecting for p-value adjusted < 0.05 in each table, one could be interested in finding which CpG islands are shared between the two groups of samples and findOverlaps is really straightforward in doing so. What is needed first is to convert the CpG

island genomic coordinates into GRanges object and give them to findOverlaps(…) as query and subject arguments:

```
## Loading the required packages
library(data.table)
library(GenomicRanges)

setwd("/Users/antoniofadda/Desktop/Tesi_PhD")

# Loading the CpG island differential methylation table
diffade <- fread("diffMethTable_region_cmp1_cpgislands_ADE.csv", data.table = F)
diffcrc <- fread("diffMethTable_region_cmp1_cpgislands_CRC.csv", data.table = F)


# Selecting CpG island for p-value adj < 0.05
diffade0.05 <- subset(diffade, diffcrc$comb.p.adj.fdr < 0.05)
diffcrc0.05 <- subset(diffcrc, diffcrc$comb.p.adj.fdr < 0.05)


head(diffade0.05)

##   Chromosome    Start      End mean.mean.ADENOMA mean.mean.CONTROLLO
## 1       chr1  1474963  1475220         0.4821670           0.1709788
## 2       chr1 13910138 13910868         0.4095413           0.1716099
## 3       chr1 14924611 14925993         0.4046099           0.1710108
## 4       chr1 16085148 16085862         0.4545077           0.1650601
## 5       chr1 35394748 35396206         0.3954926           0.1015762
## 6       chr1 37498378 37500624         0.4854369           0.2727011


##   mean.mean.diff comb.p.adj.fdr
## 1      0.3111882     -0.7751399
## 2      0.2379314     -0.5063615
## 3      0.2335991     -0.6345648
## 4      0.2894476     -0.7682760
## 5      0.2939164     -0.7298783
## 6      0.2127358     -0.7733608

head(diffcrc0.05)

##   Chromosome   Start     End mean.mean.N mean.mean.T mean.mean.diff
## 1       chr1  949330  949851   0.8143145   0.6510705      0.1632439
## 2       chr1 1011510 1013402   0.6277643   0.5218388      0.1059255
## 3       chr1 1145569 1145878   0.7745339   0.6532945      0.1212394
## 4       chr1 1267086 1267286   0.7404203   0.5659855      0.1744347
## 5       chr1 1897582 1897786   0.7481844   0.6462189      0.1019655
## 6       chr1 2066368 2066666   0.8695007   0.7582984      0.1112023
##   comb.p.adj.fdr
## 1     -0.7399810
## 2     -0.3565740
## 3     -0.6358825
## 4     -0.9049239
## 5     -0.4376749
## 6     -0.2459306

# Making GRanges object
diffade0.05_ranges <- makeGRangesFromDataFrame(diffade, keep.extra.columns = T)
diffcrc0.05_ranges <- makeGRangesFromDataFrame(diffcrc, keep.extra.columns = T)
```

```
# Finding the common Cpg islands between CRCs and adenomas
ov <- findOverlaps(diffade0.05_ranges, diffcrc0.05_ranges)
CpGIsland_common <- data.frame(diffade0.05[queryHits(ov),],
diffcrc0.05[subjectHits(ov),])


head(CpGIsland_common)

##     Chromosome     Start       End mean.mean.ADENOMA mean.mean.CONTROLLO
## 20        chr1   2983926   2987962         0.3008001           0.0827585
## 23        chr1   3111580   3111909         0.3423944           0.1260499
## 59        chr1 161008378 161008830         0.4184193           0.1975752
## 128       chr7   5886993   5887483         0.4746991           0.2667575
## 196       chr9 133805005 133805453         0.6224305           0.4186714
## 262      chr10    459693    460058         0.4431057           0.2301509


##     mean.mean.diff comb.p.adj.fdr Chromosome.1   Start.1     End.1
## 20       0.2180416     -0.5125252         chr1   2983926   2987962
## 23       0.2163445     -0.6142758         chr1   3111580   3111909
## 59       0.2208441     -0.6776491         chr1 161008378 161008830
## 128      0.2079416     -0.5407409         chr7   5886993   5887483
## 196      0.2037591     -0.7986960         chr9 133805005 133805453
## 262      0.2129548     -0.5646725        chr10    459693    460058
##     mean.mean.N mean.mean.T mean.mean.diff.1 comb.p.adj.fdr.1
## 20    0.8451489   0.7150053        0.1301437       -0.8368488
## 23    0.7911255   0.6740829        0.1170427       -0.5554766
## 59    0.2144357   0.1059864        0.1084493       -0.7967306
## 128   0.8458280   0.6896105        0.1562176       -0.9252555
## 196   0.7627675   0.5542964        0.2084711       -0.8404978
## 262   0.8876532   0.7856511        0.1020021       -0.2244568
```

To proceed with the pathways enrichment analysis these ranges must first be annotated with a gene name. The annotation of CpG islands could be done according to their proximity to other ranges (proximity criterion), such as gene structure, or to their overlap with the promoter region of a gene (functional criterion).

The proximity criterion involves the use of the FDb.InfiniumMethylation.hg19 package that merges all of the existing Illumina Infinium DNA methylation probe annotations into a FeatureDb object, a data container for storing, querying and analysing large sets of genomic annotations. The main function getNearestGene(…) takes as argument the GRanges object to be annotated and return, as the same function name suggest, the gene symbol that is much close to the CpG island of interest.

```
## Loading the required packages
library(FDb.InfiniumMethylation.hg19)

setwd("/Users/antoniofadda/Desktop/Tesi_PhD")

# Making GRanges object of common CpG island
CpGIsland_common <- CpGIsland_common[,c(1:3)]
```

```
CpGIsland_common_ranges <- makeGRangesFromDataFrame(CpGIsland_common)

# Annotating according to the proximity criterion
Gene_names <- getNearestGene(CpGIsland_common_ranges)
head(Gene_names)
```

```
##     queryHits subjectHits distance nearestGeneSymbol
## 20          1       18295        0            PRDM16
## 23          2       18295        0            PRDM16
## 59          3       14334        0              F11R
## 128         4       12766        0            ZNF815P
## 196         5       23016        0            FIBCD1
## 262         6        7377        0             DIP2C

## 601        17        5330        0          C19orf25
## 634        18       10039        0              FFAR1
## 660        19       11879        0              KCNQ2
## 666        20        6342        0              DSCAM
## 671        21       18896        0               CERK
```

The second annotation method could be defined as functional criterion given the possible repressive role of the transcription if a hypermetilated CpG Island overlap a gene promoter region. The TxDb family of Bioncoductor packages and data objects manages and stores the range of each exon, the coding range, the transcript ID, the gene ID, and metadata about the source of the transcript information. In particular TxDb.Hsapiens.UCSC.hg19.knownGene package has been used to retrieve the transcription start site (TSS) of each transcript in the human transcriptome. The promoters(…) function has been used to calculate the promoter region of each transcript creating a range of 2000 bp upstream and 1000 bp downstream around each TSS. The CpG island ranges was then overlapped to these promoter ranges to find out which CpG island overlap with.

```
## Loading the required packages
library(Homo.sapiens)
Homo.sapiens
```

```
## OrganismDb Object:
## # Includes GODb Object:  GO.db
## # With data about:  Gene Ontology
## # Includes OrgDb Object:  org.Hs.eg.db
## # Gene data about:  Homo sapiens
## # Taxonomy Id:  9606
## # Includes TxDb Object:  TxDb.Hsapiens.UCSC.hg19.knownGene
## # Transcriptome data about:  Homo sapiens
## # Based on genome:  hg19
## # The OrgDb gene id ENTREZID is mapped to the TxDb gene id GENEID .
```

```
# retriving the transcripts coordinates and calculating the promoter region
transx <- transcripts(Homo.sapiens, columns=c("TXNAME","SYMBOL"))
prom <- promoters(transx, upstream = 2000, downstream = 1000)
```

```
# Overlap between CpG island and promoters
CpG_overlpping_promoter <- subsetByOverlaps(prom, CpGIsland_common_ranges)
head(CpG_overlpping_promoter)
```

```
## GRanges object with 6 ranges and 2 metadata columns:
##       seqnames                ranges strand |          TXNAME
##          <Rle>             <IRanges>  <Rle> | <CharacterList>
##   [1]     chr1 [  2983742,   2986741]      + |      uc001akc.3
##   [2]     chr1 [  2983742,   2986741]      + |      uc001ake.3
##   [3]     chr1 [  2983742,   2986741]      + |      uc001akf.3
##   [4]     chr1 [  2983742,   2986741]      + |      uc009vlh.3
##   [5]     chr1 [  2983290,   2986289]      - |      uc010nzg.1
##   [6]     chr1 [161007775, 161010774]      - |      uc001fxf.4
##                SYMBOL
##       <CharacterList>
##   [1]          PRDM16
##   [2]          PRDM16
##   [3]          PRDM16
##   [4]          PRDM16
##   [5]       LINC00982
##   [6]            F11R
##   -------
##   seqinfo: 93 sequences (1 circular) from hg19 genome
```

There could be some inconsistencies between the two annotation methods because CpG islands do not always overlap the promoter region but could be located in other genomic regions such as gene body or 3' end. This discrepancy can easily be exemplified by the situation reported in Figure 1 in which, using the proximity criterion will be the Gene 1 that turn out to be closer to the CpG island but localized into the 3' of the gene. While with the functional criterion will be the Gene 2 to be annotated as the CpG island overlap its promoter and therefore more likely involved in its regulation. The main disadvantage of the functional criterion is that not all the CpG islands in the genome overlap with a gene promoter region (only in 60% of the genes see introduction) therefore a consistent fraction would remain excluded from the annotation especially when the number of CpG islands to be annotated increase.

Figure 1: Schematic representation of genomic coordinates

Samples for transcriptome analysis

RNA was extracted from remaining tissue available from the first methylome step using the RNeasy Lipid Tissue Mini Kit (Qiagen). After controlling for integrity, seventeen tumoral and 2 peritumoral samples achieved an optimal RIN (>7) and were processed for further analyses. In addition, 2 new peritumoral samples were recruited for the whole genome gene expression analysis, that was performed using the HumanHT-12 v4 Expression BeadChip Kit according to manufacturer's protocol. The RNA, quantified by spectrophotometric (NanoDrop) and fluorometric (PicoGreen) reading, is qualitatively checked by means of the tool Bioanalyzer2100 (Agilent Technologies), which provides an index of integrity of the RNA (RNA Integrity Number, RIN), ranging between 0 (complete degradation) and 10 (excellent quality); in our study, were further processed only the samples with RIN>7. 200 ng of RNA are then copied to cDNA and recopied to cRNA, marking them; then hybridized to the chip, colored and, subsequently to scan, the expression levels of the transcripts tested will be expressed as fluorescence intensity values.

Transcriptome analysis

After a series of steps to normalize the fluorescence intensity value obtained for each probe, differential expression analyses were carried out between CRCs and peritumoral samples. The differential expression is given in this case in fold change (FC). False Discovery Rate of 0.05 was chosen as a threshold for significance.

Samples for Real-Time qRT-PCR validation

In order to verify whether the hypermethylation of certain gene promoters in CRCs vs. peritumoral tissues, effectively causes a downregulation in gene expression, a total of 26 RNA samples (eight CRC with matched peritumoral tissues and ten individual CRCs) were tested by qRT-PCR. The same samples were used to validate the downregulation, in tumor tissues, of genes identified by the transcriptome analysis. RNA extraction from tumor and peritumoral tissues stored in RNA later was performed by using the RNeasy Lipid Tissue Mini Kit (Qiagen, Germany). Retro-transcription was performed starting from 1μg RNA/ sample using the High Capacity Kit (Applied Biosystems, Carlsbad, CA, USA). Gene

expression was assessed by quantitative RT-PCR using Express Sybr Green (Invitrogen, Paisley, UK) for each gene tested and for the endogenous TFRC.

qRT-PCR analysis

Data are expressed as mean ± standard deviation (SD) or mean ± standard error (SEM). Analysis of significance was done by t Student's test and by One-Way ANOVA using the GraphPad software (La Jolla, California). P-values were considered significant at $p< 0.05$.

Samples for pyrosequencing methylation validation

A validation of three selected CpG islands hypermethylation was performed in 78 CRCs and the respective 78 peritumoral samples. Tissue samples were collected at the Department of Surgical Sciences, University of Cagliari (Italy).

Pyrosequencing analysis

Primers were designed for three selected CpG islands using the PyroMark® Assay Design SW 2.0 (Qiagen). Amplification was carried out using the Platinum® Taq (Life technologies). PCR products were purified on the PyroMark Q24 Vacuum Workstation according to manufacturer protocol and annealed with the sequencing primer before being run on the PyroMark Q24 (Qiagen). Pyrograms were analyzed using PyroMark Q24 Software.

Stool samples for methyl-BEAMing analyses

Stool samples were collected from 24 patients with colorectal cancer and taken at the time of surgical resection. All stools sample were immediatly frozen after collection and stored at −80°C until being processed. DNA extraction was performed using the QIAamp DNA Stool Mini Kit according to the manufacturer's instructions. All samples were collected at Department of Surgical Sciences, University of Cagliari and Department of Clinical and experimental medicine, University of Sassari.

Plasma samples for methyl-BEAMing analyses

Blood draws were collected from forty five cases of CRC enrolled at the medical oncology department of the Candiolo Cancer Institute-FPO, IRCCS (Torino, Italy) between November 2015 and April 2016. Twelve cases were under adjuvant therapy after surgical resection of their lesion and were considered with no evidence of disease (NED). Remaining cases (N=33) were metastatic CRC with different level of tumor burden. Collection was approved by local ethic committee. Whole blood were processed within samples three hours after collection. Samples were centrifuged at 1600g for 10 minutes for phase separation. Plasma was collected and submitted to a second centrifugation step at 3000g for 10 minutes to remove platelets and cell debris. Upper phase was collected, aliquoted and stored at -80°C until further processing. One milliliter of plasma was processed for DNA extraction using the Maxwell® RSC ccfDNA Plasma Kit (Promega) using 100μl for the elution volume. Bisulfite conversion was performed with the EZ DNA Methylation Gold kit (Zymo Research) using 20μl of DNA as initial input, and twice 20μl for the elution.

MethylBEAMing analysis

Methyl-BEAMING is an ultra-sensitive emulsion PCR based technique that, counting the methylated and unmethylated molecules one-by-one increase the signal- to-noise ratio of the assay, with an improvement sensitivity for disease detection and test specificity. Briefly, a first amplification is carried out allowing the enrichment of the locus of interest. Amplicons of ~100 bp were chosen to accommodate the small size of circulating DNA molecules. The PCR primers were designed to amplify bisulfite-converted products derived from both methylated and unmethylated templates. PCR products were diluted and reamplified in an emulsion PCR allowing physical separation and independent amplification of the different templates. Emulsion was performed by repetitive pipetting. PCR amplification of individual DNA molecules takes place within aqueous nanocompartments suspended in a continuous oil phase. Each aqueous nanocompartment contains the DNA polymerase, cofactors and dNTPs required for PCR. When a compartment contains a single DNA template molecule as well as a bead, each bead ends up with thousands of identical copies of the template within its nanocompartment, a process similar to that resulting from cloning an individual DNA fragment into a plasmid vector to form a bacterial colony. After PCR, the beads are collected by breaking the emulsion, and their status is individually assessed by incubation with

fluorescent hybridization probes. In methyl-BEAMing, the status of harvested beads is interrogated by fluorescent probes that specifically hybridize to bisulfite-converted sequences derived from either methylated or unmethylated parental DNA sequences. Flow cytometry then provides an accurate enumeration of the original template molecules that are methylated or unmethylated within the queried sequence. Because each bead contains thousands of molecules of the identical sequence, the signal to noise ratio obtained by hybridization or enzymatic assays is extremely high and it is a quantitative method because beads obtained via BEAMing accurately reflect the DNA diversity present in template populations. To perform methyl-BEAMING assay on three selected CpG islands we used the same primers used for pyrosequencing coupled with Tag sequence as previously described [182]. Two microliters of bisulfite converted cfDNA was amplified in replicate and processed following the same protocol than previously described [188]. Purified beads were run on a BD Accuri C6 (Becton-Dickinson), methylation percentage was expressed as the number of event in the methylated gate divided by the sum of events in methylated and unmethylated gates multiply by 100.

RNAseq data and differential expression analysis

To facilitate cross-sample comparison and differential expression analysis, the Upper Quartile normalized FPKM (UQ-FPKM) values has been obtained from TCGA open-access data for 478 CRC solid tumor tissues and 41 solid normal tissues from the GDC (Genomic Data Commons) data portal. The data downloading and differential expression analysis has been conducted using the Bioconductor "TCGAbiolinks" package.

Gene Set/pathways Enrichment Analysis

Gene Set/pathways Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. This powerful approach allows to extract biological insight from a large list of genes, such as derived from methylome or transcriptome studies, to give a meaningful biological interpretation of the results, (i.e identifying groups of genes that function in the same pathways), and to reduce the number of candidate genes to a manageable number for further validation. Analyzing high-throughput molecular measurements at the pathways level is very appealing for two reasons. First,

grouping thousands of genes by the pathways they are involved reduces the complexity to just several dozen pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of differently methylated or expressed genes. The GSEA was conducted by the web portal ToppGene Suite, a one-stop online assembly of computational software tools. ToppFun is the application specifically used to perform and gene list/pathways enrichment analysis. It uses as many as 14 annotation categories including GO terms, pathways, protein–protein interactions, protein functional domains, transcription factor-binding sites, microRNAs, gene tissue expressions and literatures. One or any of the 14 annotation sources can be used for feature enrichment analyses. Each feature analysis can be adjusted based on the pvalue cut-off, the multiple testing correction method or the minimum and maximum number of genes present for each annotation type. Gene-pathway annotations were compiled by combining data from KEGG, BioCarta, BioCyc, Reactome, GenMAPP, and MSigDb. Hypergeometric distribution with Bonferroni correction is used as the standard method for determining statistical significance. ToppFun takes as input a gene list (i.e HGCN symbols), such as derived from annotating differently methylated CpG islands, and gives as result a table with columns containing the name of the pathways, the calculated q-values and the "Hit in Query List" column that contain all the genes from the input list belonging the corresponding pathway. Flexible options are provided to either download results as a tab-delimited file or display as a chart.

Biomarkers selection and validation.

Biomarker selection takes place starting from the result tables of GSEA analysis conducted using as input list the genes coming from the annotation of the significant differently methylated CpG islands of CRC and adenomas. A gene list was then created, for each table, filtering out the most significant pathways, either by Bonferroni or Benjamini & Hochberg correction (q-value adjusted < 0.05), and unifying the genes in the "Hit in Query List" column. The two gene lists were then intersected each other to obtain the significant differently methylated genes, shared between CRC and adenomas, belonging to the most altered pathways in each group. This selected group of genes were then subjected to validation to verify its feasibility as potential diagnostic biomarker either as single gene or as entire panel.

As described in the introduction, biomarker is a broad term that can be associated with many aspects of the disease. Here, the term biomarker is used as indicative of the presence of cancer in a body and therefore, unless otherwise specified, is used as synonym of diagnostic biomarker.

A first approach useful to verify the biomarker ability to distinguish between two states, (i.e tumoral vs non tumoral status), is the Unsupervised Hierarchical Clustering (UHC). Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. In order to decide which clusters should be combined a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. In UHC, which does not require to pre-specify the number of clusters to be generated as opposed to partitioning methods (e.g., k-means), relationships among objects are represented by a tree whose branch lengths reflect the degree of similarity between objects. In particular, the hierarchical dendrogram can help visualize the object relationship structure between and within clusters. Hierarchical clustering uses pairwise distance matrix between observations as clustering criteria. In R software, the euclidean distance is used by default to measure the dissimilarity between each pair of observations and the function *dist(…)* were used to calculate the distance matrix. The function *hclust(…)* were used for computing UHC using as linkage criterion, which determines the distance between sets of observations as a function of the pairwise distance matrix, the maximum or complete linkage clustering. This linkage criterion computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.

Logistic regression model is one of the most common statistical technique used to estimate the probability of predict a dichotomous outcome, such as the presence or absence of a disease, given biomarker values. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = presence/absence of a disease) and a set of independent (predictor or explanatory) variables, i.e the methylation level of the single biomarker. Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of

presence of the characteristic of interest. The ability of the model to predicts dichotomous outcomes, or in other words, the ability of the biomarker to distinguish between those patients having and those not having the disease, is evaluated using two indices: sensitivity and specificity. The biomarker sensitivity describes the proportion of patients with disease that are correctly identified as such (true positive rate), whereas the biomarker specificity describes the proportion of patients without disease that are correctly identified as such (true negative result). The ideal biomarker would show 100% sensitivity and 100% specificity. In other words, the biomarker test is never positive for a disease-free patient and never negative for a patient with disease. However, this ideal scenario is rarely achieved. The biomarker sensitivity is a composite of the marker prevalence in the tumour, the efficiency of transfer of the marker to the remote media being tested, and the analytical sensitivity of the assay. The biomarker sensitivity is also enhanced by the application of panels of multiple biomarkers, because it take into account the heterogeneity of the cancer, providing more diagnostic information and cancer specificity than single-marker assays. It is also important to note that a biomarker with a sensitivity of 50% and a specificity of 50% is no better than tossing a coin to decide if the patient is harboring the disease or is disease-free. Some molecular methods are able to produce data as a categorical variable (presence or absence of methylation for each samples for a particular biomarker). For this type of measurement, the performance of a biomarker can be described simply by its sensitivity and specificity. However other technologies, such as 450k B

eadchip, are able to quantify the methylation level for each samples for a particular biomarker, generating quantitatively accurate data as a continuous variable. This data can then be dichotomized at a threshold to determine the sensitivity and specificity. However, the choice of threshold will affect these two variables differentially. Setting a higher threshold for a cancer-specific marker will increase the specificity of the biomarker, but reduce its sensitivity. Therefore, methods have been developed to describe the performance of a biomarker that take both measurements into account simultaneously. The receiver operating characteristic (ROC) curve is a fundamental tool for diagnostic test or biomarker evaluation and visually displays the interdependency of specificity and sensitivity. This curve can be plotted as the true positive rate (sensitivity; y-axis) against the false positive rate (1-specificity; x-axis), in which each point in the curve represents the fraction of cancer cases with a biomarker measurement above a threshold (true-positive rate for that threshold) versus the corresponding fraction of control subjects above the same threshold (false-positive rate for that threshold). The area under the curve (AUC) is equal to the probability that a

classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In other words, for a well performing diagnostic test or biomarker the curve is located towards the upper left corner. On the other hand a less well-performing diagnostic test or biomarker is characterized by a curve close to a diagonal line, representing a state in which sensitivity and specificity are similar. This value is a useful way to describe the performance of a biomarker with a continuous output variable, regardless of the threshold level, and is identical to the non-parametric Wilcoxon statistic. It is desirable to achieve values for sensitivity and specificity as high as possible. However, for some tests it might be acceptable to achieve a higher sensitivity by sacrificing assay specificity or vice versa. Acceptable values for sensitivity and specificity of a testing procedure can be determined by comparing to existing values of a test currently considered as gold standard. We evaluate the association of each biomarker with a binary outcome fitting a logistic regression model to the individual biomarker data and the new probabilities used to calculate the AUC, specificity and sensitivity values using the "*OptimalCutpoints*" package.

Occasionally, building a logistic regression model could be possible run into the problem of so-called complete separation and this happen when the outcome variable can be perfectly predicted by one predictor variable or a combination of predictor variables. This problem often arise with small data sets, when the event is rare or when the predictor variables are too many. The more predictor variables are in the model, the more likely separation is to occur. A solution to this problem could be the usage of a penalized logistic regression model that, adding a penalty to control properties of the regression coefficients, avoid the overfitting of the logistc regression model. This penalty causes the regression coefficients to shrink toward zero. If the shrinkage is large enough, some regression coefficients are set to zero exactly. Thus, penalized regression methods perform variable selection and coefficient estimation simultaneously. The main goal of this method is to find the simplest model, with fewer predictor variables, that also has good predictive performance from among the possible alternative models.

However one might be interested in evaluating the performance of the model using all the predictive variables together without performing a variable selection. A possible solution is to consider an alternative model. The support vector machine (SVM) is becoming popular in a wide variety of biological applications and it is a relatively new classification or prediction method developed by Cortes and Vapnik [189] in the 1990s. SVM tries to classify cases by finding a separating boundary called hyperplane. The main advantage of the SVM is that it

can, with relative ease, overcome 'the high dimensionality problem', i.e., the problem that arises when there is a large number of input variables relative to the number of available observations. SVM is one of the most well-known supervised machine learning algorithms for classification. For a given set of training data, each marked as belonging to one of two categories, SVM training algorithm develops a model by finding a hyperplane, which classifies the given data as correctly as possible by maximizing the distance between two data clusters. In practice, however, it is frequently not possible to clearly separate the given data set because some of the data points in the two classes might fall into gray area that is not easy to separate linearly. As one of the solutions for this problem, data are mapped to a higher dimension such that the two classes could be separable in the higher dimension called kernel function.

We evaluate the ability of the entire biomarkers panel to correctly classify between two possible conditions (i.e Tumoral vs non tumoral) building a SVM learning model randomly splitting the TCGA dataset in a training set (70%) and test set (30%) and performed a 1000 iterations to calculate the accuracy and stability of prediction performance in the training set based on AUC value. The prediction ability of the model was also evaluated on several independent in silico datasets.

In silico validation datasets

Three cross-validation datasets were retrieved from the database NCBI Gene Expression Omnibus (GEO) portal (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE48684, GSE52270, GSE53051. Processed data were used for all datasets above mentioned. The Cancer Genome Atlas (TCGA) dataset is available online at RnBeads website under Methylome Resources (http://RnBeads.mpi-inf.mpg.de/methylomes.php). For each of these datasets the mean methylation value for each CpG island of interest has been calculated and visualized by heatmap using Bioconductor package "*ComplexHeatmap*".

Microsatellite instability analysis

The microsatellite instability analysis was conducted by using the "MSI Analysis System, Version 1.2" (Promega Italia). Amplification were carried out according to manufacturer's protocol. The PCR products were then run on an ABI PRISM® 3100 - Applied Biosystems®

Genetic Analyzer and the output data were analyzed with GeneMapper® software (Applied Biosystems).

Genetic mutations screening

The genetic mutation screening in adenomas was conducted as previously reported [190, 191]. To search for KRAS mutations in CRCs, we amplified two fragments corresponding to the exons 2 and 3 (codon 1-97) (annealing 60 °C). To amplify exon 2, we used the following primers (forward: 5'-ACTGGTGGAGTATTTGATAGTGTAT-3'; Reverse: 5'-AGAATGGTCCTGCACCAGTAA-3'); Exon 3: Forward: 5'-TCCAGACTGTGTTTCTCCCT-3'; Reverse: 5'-AACCCACCTATAATGGTGAATATCT-3'). PCR products were amplified with High-Fidelity Taq polymerase (Platinum® Taq DNA Polymerase High Fidelity, Invitrogen), purified (by exonuclease 1 and shrimp alkaline phosphatase) and sequenced by fluorescent based Sanger's direct sequencing in an ABI 3130 DNA capillary sequencer.

CIMP phenotype definition

To infer a CIMP phenotype definition from the output data obtained by using Infinium® HumanMethylation450 BeadChip in the 18 CRCs and 21 adenomas, we referred to the SALSA® MLPA® kit (MRC-Holland, Amsterdam, the Netherlands) as reference panel. This ME042-C1 CIMP MS-MLPA probemix contains 31 MS-MLPA probes which detect the methylation status of promoter regions of the following 8 genes: *CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3* and *SOCS1*. In the present study, several CpG loci were analyzed for each of these gene promoters, mapping inside the MS-MLPA probe sequence. All positions/probes were scored using two different thresholds of methylation (≥ 20% or ≥ 30%), for defining a specific position/probe as methylated. To define a gene as methylated, at least one probe/position has to be methylated. To assign the CIMP positivity to a sample, at least 60% of the genes in the panel must be labelled as CIMP positive.

Results

Figure 2a shows the results of the differential methylation analysis conducted on 18 CRCs and 4 matched peritumoral samples. Red dots represent the significant differential methylated regions (p value adjusted < 0.05) in each sub-categories taken into account by the default differential methylation analysis in *RnBeads*. What can be observed is a hypomethylation of the tumoral sample in the genome-wide tiling regions that switch gradually towards an hypermethylation as the focus of the differential methylation analysis is restricted to the coding region (genes) and even more to the regulatory regions such as promoters and CpG islands. Focusing on the CpG islands results, which show a more pronounced hypermethylation in the tumoral samples compared to the other regions, the number of CpG islands significantly altered in CRC (p-value adjusted < 0.05) were 875. The annotation of the CpG islands based on 450k manifest using a proximity criterion, which consist in finding the closest gene to each CpG island, has allowed to create a gene list whose CpG islands were significantly altered by aberrant methylation. This gene list was then subjected to GSEA using the bioinformatics ToppGene Suite. This analysis allowed the identification of the pathways most affected by aberrant methylation (Table 3). After narrowing the focus to the pathways significantly involved by stringent statistical corrections, the 10 pathways mostly affected by gene promoter methylation status alterations were: Wnt signaling, Neuronal System, Cadherin signaling, Transmission across Chemical Synapses, Neuroactive ligand-receptor interaction, Neurotransmitter Release Cycle, GABAergic synapse, Core extracellular matrix, Calcium signaling, Cholinergic synapse.

Figure 2b also show the results of the differential methylation analysis conducted on 21 adenomas *vs*. 3 samples of control mucosa. The observed methylation pattern, a genome wide hypomethylation and site specific CpG islands hypermethylation, is similar to that observed in CRCs. The number of CpG islands significantly altered in adenomas (p-value < 0.05) was 2393. The number of CpG islands that resist to the multiple test correction were too low to guarantee robusteness to the subsequent analysis therefore we used the nominal p-value event though this increase the probability of making type I error (false positive). However the GSEA analysis show that the most affected pathways (Table 4) were largely comparable to those identified in CRCs. Filtering out the most significant pathways, either by

Bonferroni or Benjamini & Hochberg correction (q-value adjusted < 0.05) the 10 pathways mostly affected by gene promoter methylation status alterations were: Cadherin signaling pathway, Wnt signaling pathway, Neuroactive ligand-receptor interaction, Neuronal System, Neural Crest Differentiation, Extracellular matrix organization, Transmission across Chemical Synapses, Calcium signaling pathway, GPCR ligand binding, Nicotine addiction.

Comparing the GSEA results of CRCs and adenomas, it appeared that the two groups of samples shared the most significantly modified pathways. Alongside to pathways that are not perviously described as associated with CRCs, such as Wnt signaling and Cadherin signaling, there are others that are not previously described as associated with CRC: Neuronal System, Transmission across Chemical Synapses, Neuroactive ligand-receptor interaction (Figure 2c). Of note, most of the genes whose associated CpG islands were altered in CRCs, were already aberrant in adenomas. From the intersection of the gene lists belonging to the significantly altered pathways in CRC and adenomas ,we selected 74 genes whose altered CpG island are shared between CRC and adenomas. In Table 5 are listed 74 CpG islands resulting from the described intersection, with the respective average beta values and Δs calculated between the average value in the tested samples (CRCs or adenomas) and in the respective controls.

The UHC analysis (Figure 3) conducted on these 74 CpG islands show that all CRCs, except one sample (352T), clusterize together well separated from the peritoumoral counterpart. The UHC on the same CpG islands in adenomas (Figure 4) reveal the presence of two different clusters: a cluster made up of the samples whose methylation pattern resemble to the normal mucosa and the other one whose methylation pattern is closely related to the one of CRCs samples. Of note two islands behaved in an opposite way resulting highly methylated in peritumoral samples both in CRC and adenomas. No correlation trend was observed between methylation pattern and staging, localization or mutational pattern in both CRCs and adenomas.

This panel of 74 altered CpG island is therefore able to clearly distinguish CRC and peritumoral counterpart. These alterations are also detectable in adenomas suggesting that these 74 CpG islands could be used as biomarker panel for early detection of colorectal cancer.

To verify the robustness of our results, the biomarker panel was validated *in silico* using publicly available dataset coming from studies performed using 450k technology. First we

tested the ability of our biomarkers to clearly distinguish CRC and peritumoral counterpart using the methylation data from two databases:

- 248 colon and 94 rectum adenocarcinomas, vs. 37 colon and 7 rectum normal tissues from The Cancer Genome Atlas (TCGA) consortium;

- GSE48684, consisting of samples from 42 adenomas, 64 carcinomas, and 41 normal mucosa from colon [195].

The UHC analysis on the TCGA dataset (Figures 5) and GSE48684 (Figures 6) show that the panel is capable of clusterize the majority of CRCs and adenomas from the peritumoral/ normal counterpart. The performance of each single biomarker was evaluated from an analytical point of view, fitting a logistic regression model to the individual biomarker data, using the TCGA dataset as reference. As shown in Table 6, the specificity of many markers was equal to 1, i.e. 100% (ranging from 0.89 and 1), the sensitivity was ≥ 0.9 in over 70% of the islands (ranging from 0.7 and 0.97) and AUC value above the 0.9 for most of the biomarkers. We also evaluated the specificity, sensitivity and AUC for the entire biomarkers panel fitting a SVM model, obtaining respectively: SP = 1; SE = 0.9992; AUC = 0.9999.

Afterwards, the biomarkers specificity for CRCs was evaluated examining the GSE52270 dataset [195], consisting of 103 CRC samples, 18 colon peritumoral tissues, 66 breast cancer and 19 no-tumoral breast, 48 glioblastomas and 10 white matter samples. The UHC analysis on GSE52270 (Figure 7) shows that the hypermethylation pattern represent a distinctive trait of CRCs and colorectal adenomas with rare exception only for few biomarkers. Of note, the only two CpG islands that behave in opposite way resulting hypomethylated in CRCs and adenomas, show this characteristic tendency to the hypermethylation in all samples except for CRCs and adenomas. The second dataset, (GSE53051) [196], is similar to the previous one but include colon cancer metastases localized in the lung and in the liver (indicated in Figure 8 with longer reddish barline). The UHC (Figure 8) shows that colon cancer metastases clusterize along with CRCs, with the exception of two cases.

Even though we have validated our biomarkers as predictor of colon cancer using a huge amount of *in silico* data, to propose their use in clinical practice, such as in patients screening, the next step was to test the methylation alteration of some selected CpG islands in DNA extracted from more accessible biological matrices, such as stool samples and plasma. The CpG islands were selected based on both a large differential methylation between tumor and non-tumor tissue and the feasibility of the assay design. The three

selected genes, as the rest of the panel, are protected by patent therefore at the moment we can not show their names but are indicated as biomarker 1, biomarker 2 and biomarker 3. Since the tumoral DNA, extracted from such biological matrices, is subjected to a rapid degradation the assay was design so that significant probes were not distant more than 150 bp. Before to proceed with the analysis on stool and plasma samples we further verify the strength of the three selected markers performing a methylation analysis by pyrosequencing in a second data set of 78 tumoral and 78 matched peritumoral. All three selected CpG islands were significantly hypermethylated in tumor vs. peritumoral samples (Figure 9).

Methylation of the three selected biomarkers was assessed by methyl-BEAMING in DNA isolated from stool samples of patients who were diagnosed with colorectal cancer, taken at the time of surgical resection. As shown in Figure 10, all except three samples tested (87.5%) showed more than 1% of methylation for at least one of the three markers. In particular, 79.2% of samples showed more than 1% of methylation at biomarker 1 (average percentage of methylation equal to 21%); 70.8% of samples at biomarker 2 (methylation average 10%); 62.5% of samples at biomarker 3 (methylation average 13%).

Methylation of the three selected biomarkers was then assessed by methyl-BEAMING in DNA isolated from plasma samples of 45 colorectal cancer patients divided into two cohorts:

- 12 patients were under adjuvant therapy after surgical resection of their lesion and were considered with no evidence of disease (NED).

- 33 colorectal cancer metastatic patients with different level of tumor burden

the Assay detected the presence of tumor DNA in at least one replicate for 43, 45 and 41 cases for biomarker 1, biomarker 2 and biomarker 3 respectively. The methylation level was significantly different for biomarker 1 and biomarker 2 between the NED patients and the metastatic ones (u-test, biomarker 1: p value =0.029; biomarker 2: p value =0.024). This result suggest the specificity of the two selected biomakers for colorectal cancer patients. Biomarker 3 did not show any difference, possibly due to low number of cases which displayed methylation (Figure 11). Looking at the clinical features, the metastatic samples were stratified in two subgroups (CEA-low and CEA-high) based on CEA level threshold of 5 ng/dl. CEA (which stand for Carcinoembryonic antigen) is a glycoprotein and its level increased in serum during cancer progression. It is a conventional cancer marker used to monitor disease progression, cancer relapse, and response to therapy in patients with CRC (prognostic biomarker). Dividing the sample in the three groups (NED, CEA-low, CEA-

high) there were no significant differences between NED and CEA-low while significantly higher levels of methylation were evident for biomarker 1 and 2 in the CEA-high group. Absence of alteration observed in the low-CEA subgroup might be due to specific biologic features of these tumors, such as early tumor stage and absence of vascular network, that impair the release of circulating markers. Moreover, the current assay was designed for monitoring purposes and not early detection as expected in clinical field. Therefore assay improvements for the early detection are suggested in a possible future clinical application.

To investigate whether the alterations found in the CRC methylome result in dysregulation of gene expression, we performed the analysis of the transcriptome using the Illumina Whole Genome Gene Expression technology in the same tumoral and peritumoral samples used for the methylome analysis. As expected, the UHC analysis revealed that all CRCs clustered well separated from peritumoral samples (Figure 12). The results of the differential expression analysis showed the presence of 725 downregulated transcripts and 381 upregulated transcripts. The GSEA analysis conducted on the disregulated genes didn't show pathways directly overlapping with those coming from the methylome analysis, except for those relating to the extracellular matrix. However a number of downregulated pathways, namely those involved in the amines degradation, are likely downstream of those altered as a consequences of hypermethylation (Figure 13). Of note, none of the 74 biomarker genes resulted dysregulated by the transcriptome analysis but displayed an extremely low level of expression close to the background intensity level of the Beadchip. Therefore to test whether the alterations found in these CpG islands could modulate the expression of their target genes, the expression level of three selected genes was determined by qRT-PCR on 8 CRCs and their matched peritumoral tissues and 10 additional CRCs. As shown, all the three tested genes were strongly down-regulated in cancer (Figure 14).

Since it was not possible to perform a qRT-PCR analysis for all 74 CpG islands, we validated *in silico* using the RNA seq data from the TCGA samples. As shown in Figure 15, almost all the selected altered CpG islands were located in the promoter region (defined as the sequence between 2kb upstream and 1kb downstream the TSS), and the majority of them (> 70%) were associated to a down-regulation of the corresponding genes in the tumor tissues. There were only three CpG islands that appear strongly upregulated; actually they were not located into the related promoter gene regions but instead in the gene body or downstream.

Discussion and conclusions

The results of this study have identified a panel of early biomarkers for CRC.

The analysis of the CRC methylome revealed alterations in the methylation status of CpG loci compared to peritumoral tissues. In particular, epimutations characterizing CRC frequently relate to CpG islands, typically occurring at or near the transcription start site of genes, so likely involved in the regulation of gene expression. The vast majority of these epimutations are hypermethylation in CRC respect to peritumoral tissue. As expected from literature, analysis of CpG loci scattered along the genome, but not localized in CpG islands, are rather more hypomethylated in CRC. The GSEA analysis of genes likely dysregulated by the epimutations identified, has shown the involvement of pathways in part not yet described as associated with CRC. The data suggest that the mechanisms most affected by hypermethylation, concern the reception and transmission of nerve impulses through neuroactive ligands, in addition to Wnt signaling and Cadherin signaling pathways. As shown in Figures 13 and 14, the majority of genes whose CpG islands are hypermethylated, are effectively silenced, at the mRNA level, both in the cases analyzed in wet in laboratory (for some genes) (Figure 13), and in silico in the TCGA series (Figure 14). Interpreting what can be deduced from this data, it appears that, to become cancerous, the cell isolates itself, closing ways of signal reception, since the early stages of neoplastic transformation.

As known, epigenetic alterations probably represent the first modification that the cell undergoes in the carcinogenetic process. In order to verify whether the epimutations identified in CRC in this study are important to the onset of cancer and therefore already present at the stage of adenoma, we have extended the methylome analysis to a group of adenomas, comparing them to control mucosa. The results are completely in line with the premises: also adenomas are characterized by the presence of hypermethylated CpG islands, surrounded by a broader genomic context widely hypomethylated; the gene ontology analysis showed that the most altered pathways largely correspond to those identified in CRC; genes altered in the promoter, belonging to these pathways, are largely shared by CRC and adenomas. Performing a cross comparison analysis between CRC and adenomas methylation alterations, it was possible to select a panel of 74 CpG islands, which we define therefore early biomarkers of the CRC carcinogenic process. To test the strength of this panel in identifying alterations typical of CRC, since the early stages, we performed a cross

validation in silico, by examining the methylation pattern of the 74 CpG islands in other databases, including both samples from CRC, adenomas and normal mucosa from colon, both from other types of tumor tissues. The panel appear very robust and informative (sensitivity 0.9992), specific for colorectal cancer (specificity 1) and is a very good marker of early and metastatic stages. Multiple studies have investigated the use of single or combined DNA methylation-based biomarkers for diagnostic purposes[192]. Some of them have compared their result with the two only commercially available methylation biomakers: SEPT9 and VIM. Essentially all of these studies show that the SEPT9 and VIM performance ability to distinguish CRC/adenomas from the normal counterpart greatly vary between these different studies, depending on the different experimental design, but in general demonstrate the value of combining markers into a panel, because it could improve the diagnostic accuracy and achieve the highest clinical sensitivity compared to the use of single markers. Our results are not only consistent with this observation but also the performance ability of most of our single biomarkers greatly outperform those ones commercially available, based on data collected on TCGA dataset, the most powerful in silico dataset that we used in this study to validate our biomarkers.

The heterogeneity that occurs between adenomas relatively to the selected CpG islands panel, with adenomas which have a pattern much similar to normal mucosa and others definitely similar to CRC, and the lack of correlation between adenomas staging (histology) and methylation alteration, might suggest that the methylome alteration occur during the early stages but characterizes only a cohort of patients. This would therefore represent a signature, typical of carcinomas, probably defining adenomas that would follow such a destiny.

CRC is known as a “silent disease,” as many people do not have complaint until the disease is difficult to cure. Therefore, detection of patients at early stage of precancerous colorectal lesions can play a pivotal role in improving the outcome of patients. The currently available screening methods are colonoscopy and FOBT. Colonoscopy is a gold-standard screening test to identify and remove the lesion; however, its application can be limited by its invasive nature and high cost and is not routinely recommended to all risk-eligible patients, while limitations of FOBT screening includes its low sensitivity for polyps, a relatively low specificity, and false positive screens. Carcinoembryonic antigen (CEA) is a glycoprotein that its level increased in serum during cancer progression. It is a useful marker in detecting of recurrence of cancer following surgical/medical treatment. Due to low sensitivity and

specificity of this biomarker (30–40 % and 87 % respectively), it is not a suitable screening test for the early diagnosis of CRC. No conventional methods meet all of the desired criteria of an ideal screening tool, therefore, there is an urgent need to develop simple and less invasive tests with high sensitivity and specificity.

Therefore, a very ambitious goal of this study was to identify CRC cancer biomarkers to be used in large screenings and designed to detect the presence of CRC with non-invasive methods. For this reason, some of the CpG islands identified in this study as CRC biomarkers, have been sought in ctDNA of patients suffering from CRC. This test was absolutely suitable for the identification of ctDNA in the blood of patients, confirming the state of hypermethylation, thus resulting a good diagnostic test.

But even more ambitiously this study aims to find selected biomarkers in stool samples, to enhance even more their predictability. The same markers used to search for ctDNA were then tested on stool samples collected from the intestine of CRC patients during surgical resection. Carmona et al. [193] aimed to identify a set of stool-based DNA methylation markers that are suitable for early diagnosis of CRC. They selected, from a comprehensive analysis of DNA methylation profile differences in pairs of tumor and matching normal mucosa samples, three candidate markers (AGTR1, WNT2, SLIT2) and then performed a validation in stool DNA samples from CRC patients. AGTR, WNT2 and SLIT2 had a sensitivity of 21% (n = 68), 40% (n = 52) and 52% (n = 71) respectively. A panel of these genes obtained a sensitivity of 78% (n = 64) based on the criteria that at least one of the genes was methylated. By contrast, VIM and SEPT9 only yielded a sensitivity of 55 (n = 33) and 20% (n = 35) respectively. In comparison, our three selected biomarkers perfomerd better in terms of sensitivity with a percentage of samples that showed more than 1% of methylation that ranges from 62.5% to 79.2% for the three selected biomarkers. The panel of our selected genes obtained an overall sensitivity of 87,5% with all except three samples tested that showed more than 1% of methylation for at least one of the three markers. We can conclude that even in these samples the CpG islands tested came out as excellent tumor markers despite the technical difficulties to detect trace amounts of target methylated DNA among large amounts of background DNA in a biological challenging matrix such as stool.

The whole genome gene expression (WGGE) analysis conducted on the same CRC samples showed a picture of disregulation very noticeable compared to peritumoral tissues. In this analysis it was not possible to verify the expression levels of the transcripts directly silenced by hypermethylation, since these probes have shown intensity signals comparable to

the background (both in tumoral and peritumoral samples). However, bioinformatics analysis conducted on the genes significantly differentially expressed, showed a functional link between the results obtained by WGGE analysis and the study of the methylome. In fact, the pathways involved in the degradation of the main neuroactive ligands (dopamine, serotonin, norepinephrine, GABA) are strongly downregulated.

It should be noted that most of the genes found highly downregulated in this study, do not have a hypermethylated promoter and that all the 74 CpG islands coming from the methylome analysis displayed an extremely low level of expression close to the background intensity level of the Beadchip despite the qRT-PCR on three selected CpG islands, an more comprehensive in silico with RNA seq data, clearly show a strongly downregulation of the corresponding genes in the tumoral samples. Our results are consistent with the general understanding that is emerging in recent years mainly thanks to the technological progress made in the genome wide methylation analysis. A growing number of evidences suggest that transcriptionally repressed genes are also the predominant target of cancer-associated aberrant hypermethylation[51, 52]. These recent findings from the study of cancer methylomes draw parallels with our understanding of DNA methylation during normal development which does not initiate the silencing of genes but is rather a secondary or even tertiary effector and there is a complex interplay between all the elements involved in the transcription machinery. This new framework could led to a better understanding of the impact of aberrant DNA methylation pattern in carcinogenesis, to determine why specific genes are prone to targeted de novo hypermethylation while others are protected against it, to get new insights regarding the biology of the tumors.

In conclusion, this study show the relevance of multi-omics profiling on matched tissue and non-invasive cohorts along with matched cohorts of adenoma to carcinoma as indispensable approach to concurrently stratify CRC and find novel, robust biomarkers. The results of this study have shown pathways altered since the early stages of CRC carcinogenesis, so far not associated with CRC development and related to the reception and transmission of nerve impulses through neuroactive ligands. This allowed us to define, and validate in vitro and in silico, a panel of CpG islands with altered methylation both in adenomas, CRC and in colon metastases. We also demonstrated the functional relevance of these alterations in gene expression. Finally, it was shown the power of these alterations as tumoral markers traceable even in ctDNA and stool samples.

New innovative technologies have the potential to yield automated high throughput platforms for efficient performance of molecular screening assays in the future. Such systems could allow assay of multi-marker panels at high capacity and low cost. The use of a biotechnological platform allowing to reveal the methylation status of the entire CpG islands panel recommended in the present study, will be likely very useful both as CRC diagnostic test, prevention and tracking minimal residual disease.

Figures

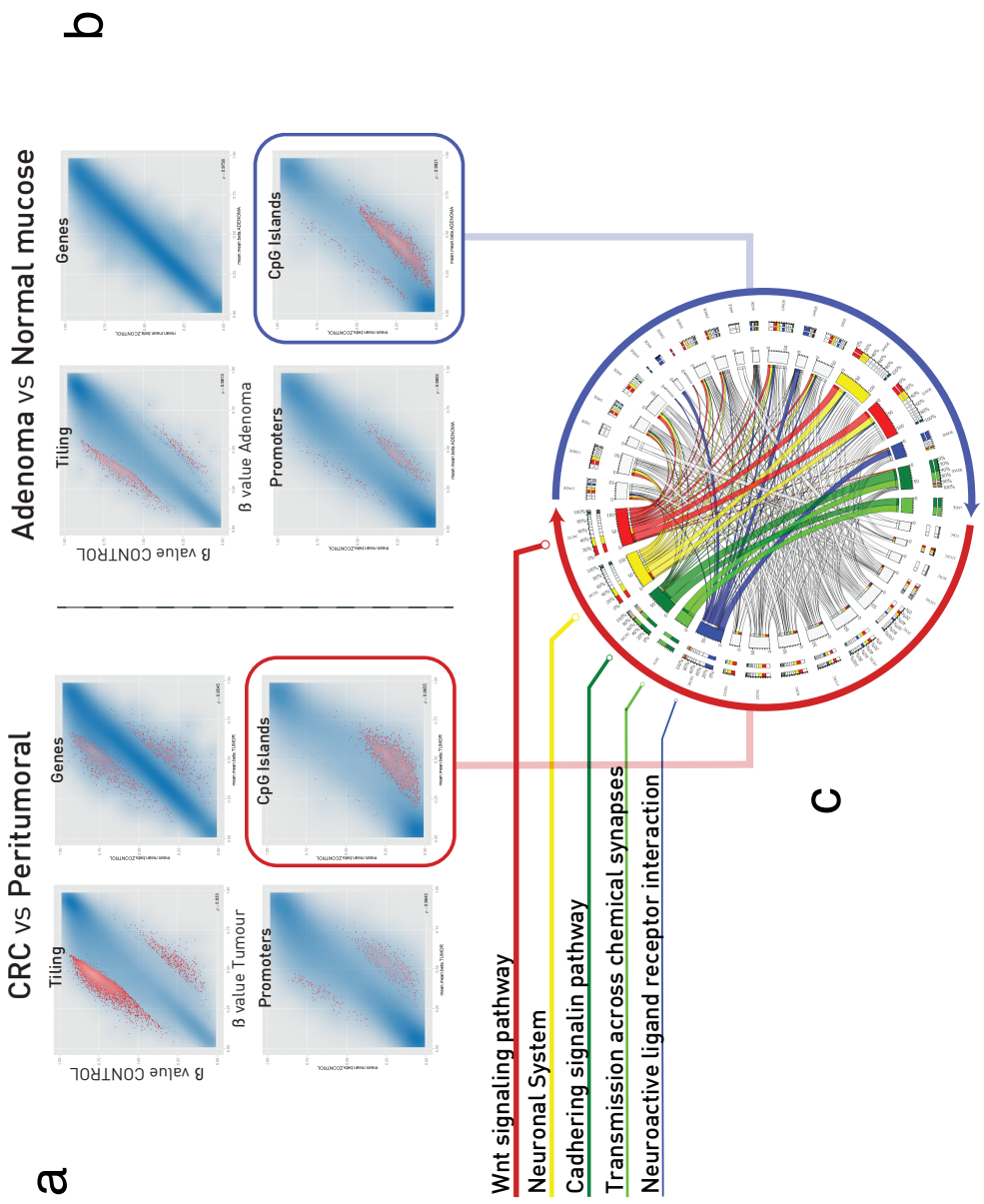

Figure 2: RnBeads differential methylation analysis performed for single CpGs and for sets of pre-defined genomic regions such as genome-wide 5kb tiling regions, genes, promoters and CpG islands, in CRC vs. peritumoral tissue samples (a) and adenomas vs. normal mucosa samples (b). Circos plot resulting by comparison of pathways significantly altered in CRC (left) and adenomas (right) (c). The red arrow indicates the pathway increased significant enrichment of altered loci (decreasing p value) in CRC; the blue arrow in adenomas. The name of the five pathways most significantly altered in CRC are given and highlighted in color. Colored beams link the shared genes.

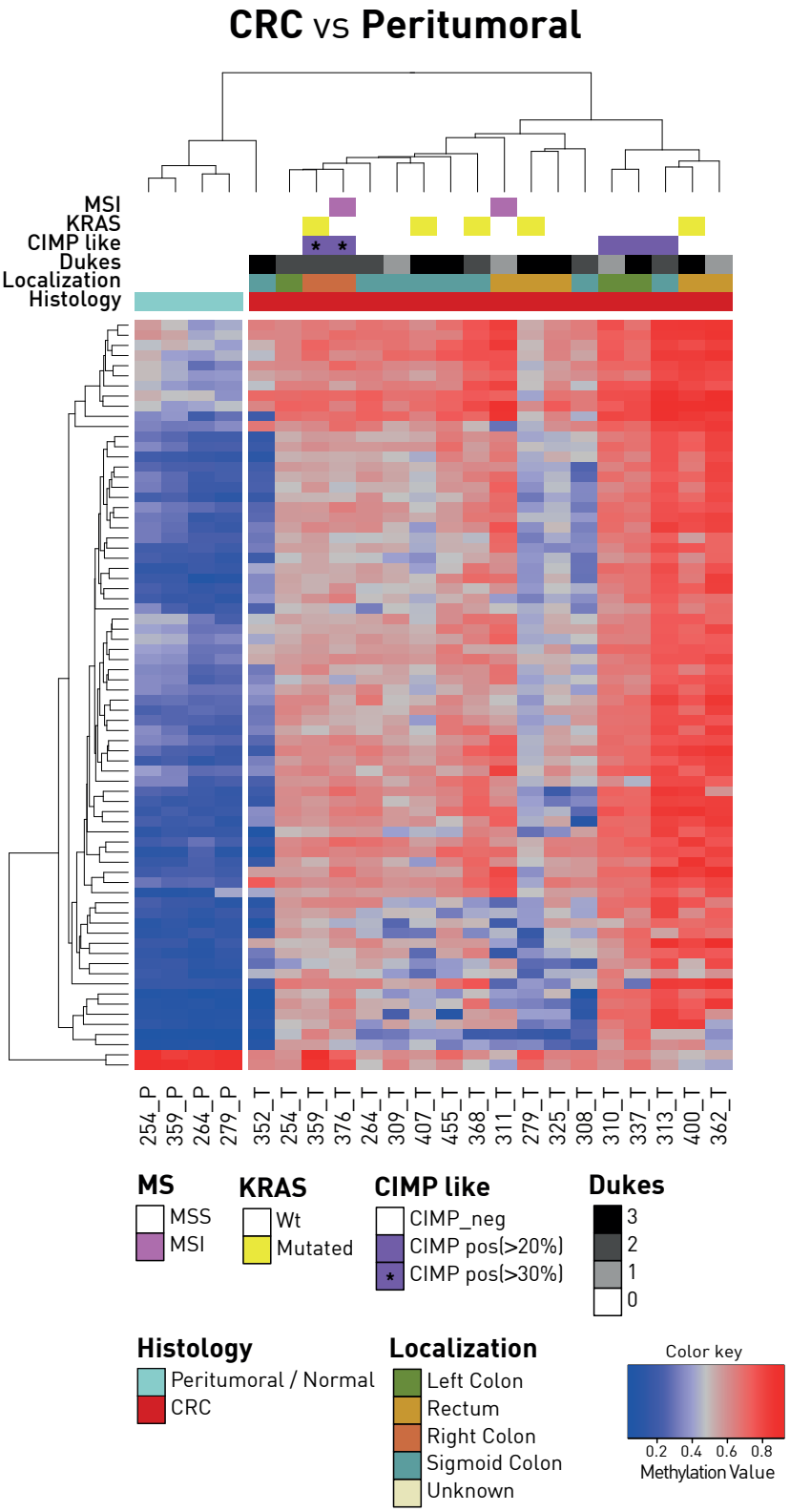

Figure 3: Heat map obtained by an unsupervised hierarchical clustering analysis performed on CRC versus peritumoral tissue samples

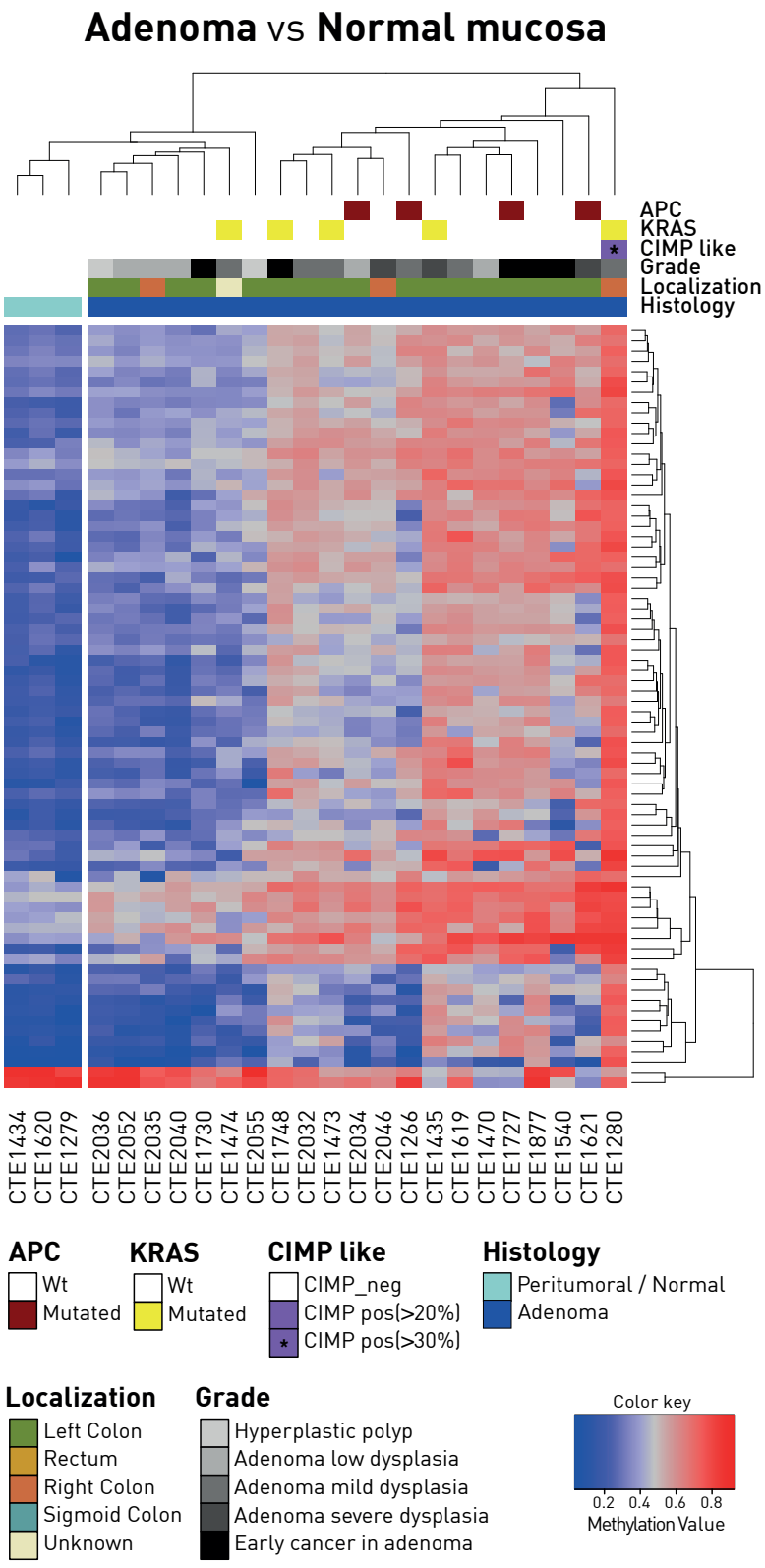

Figure 4: Heat map obtained by an unsupervised hierarchical clustering analysis performed on adenomas versus normal mucosa

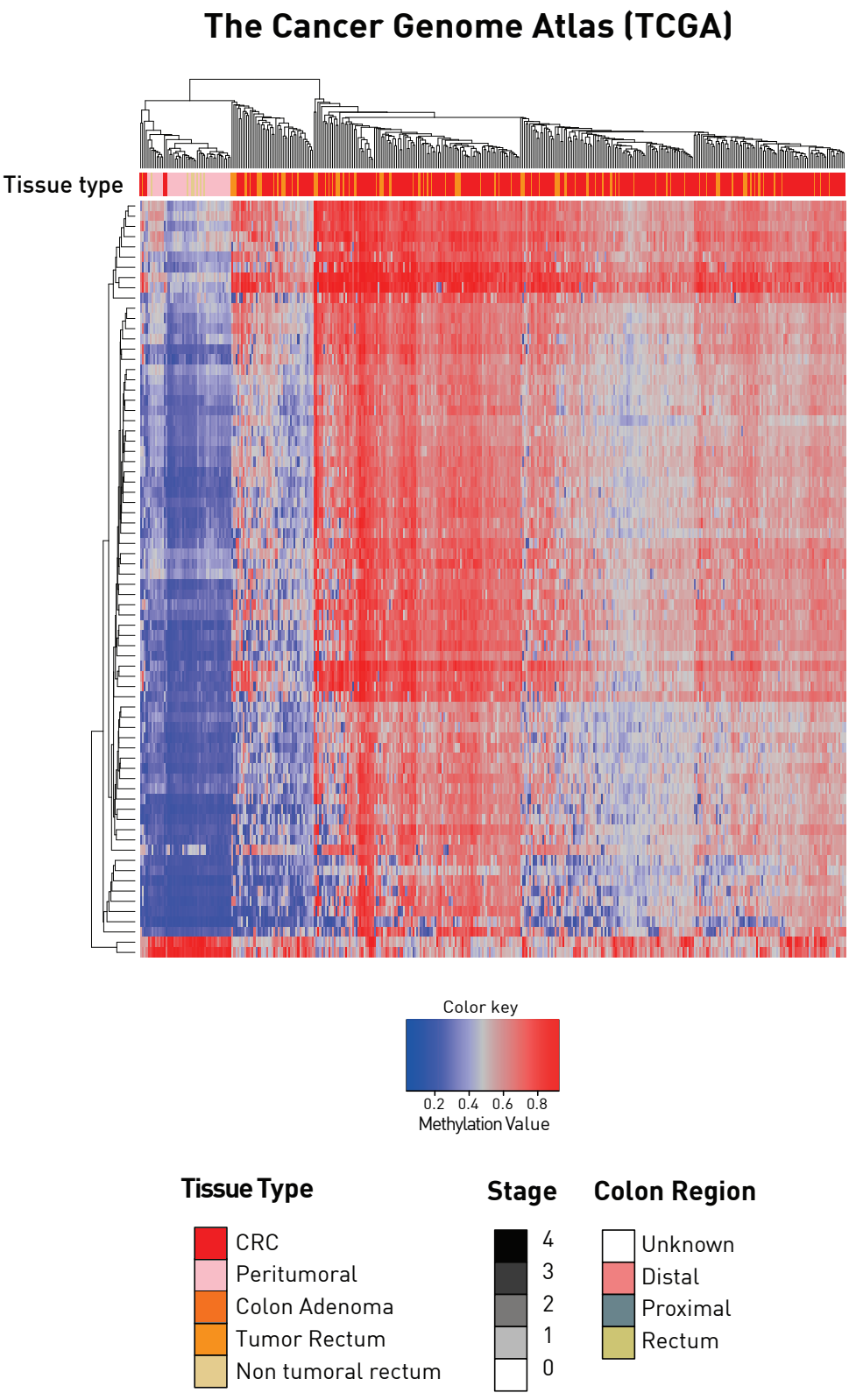

Figure 5: Heat map obtained by an unsupervised hierarchical clustering analysis performed on the TCGA data set by using the average methylation beta value for each of the 74 CpG islands.

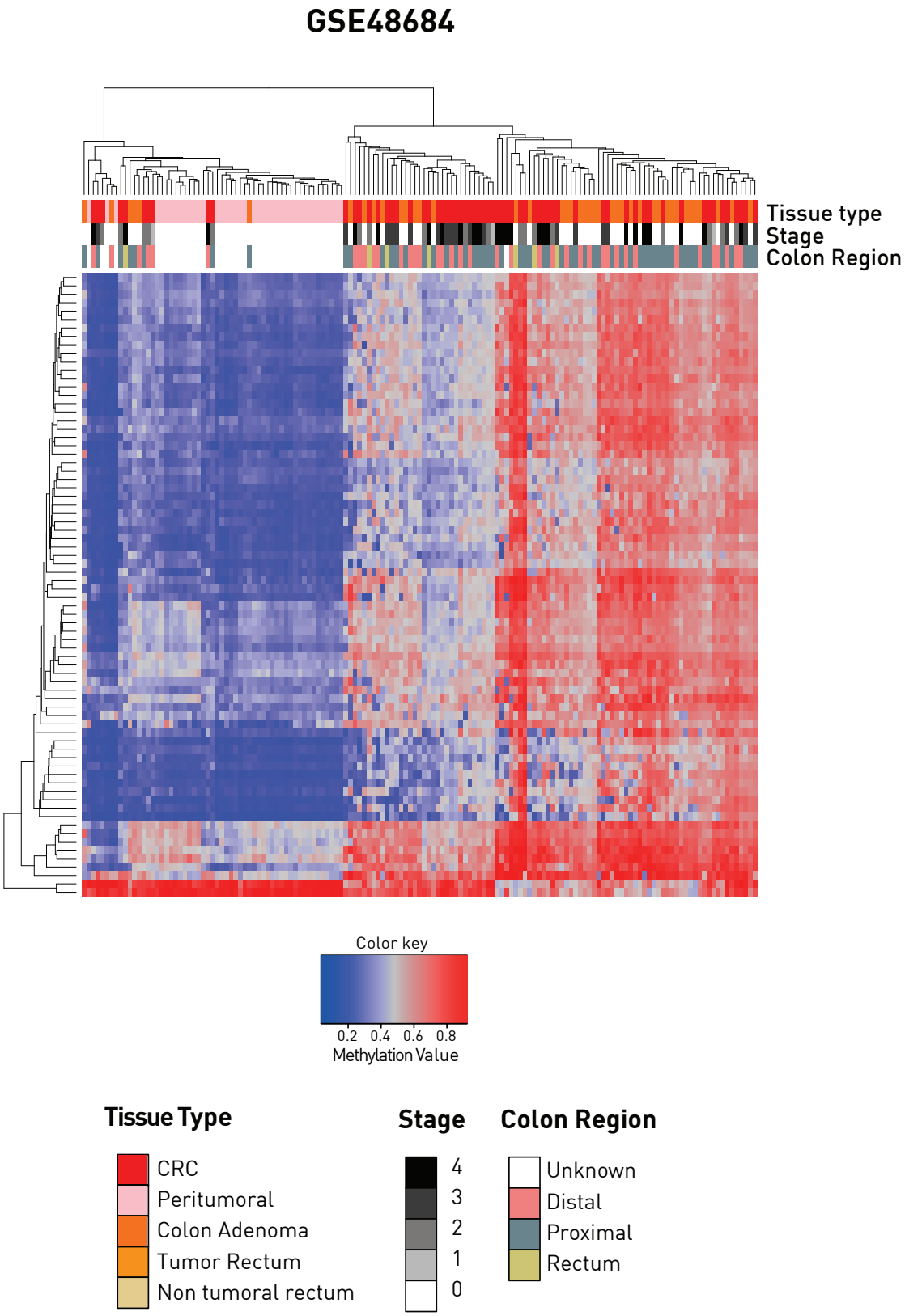

Figure 6: Heat map obtained by an unsupervised hierarchical clustering analysis performed on the GSE48684 data set by using the average methylation beta value for each of the 74 CpG islands.

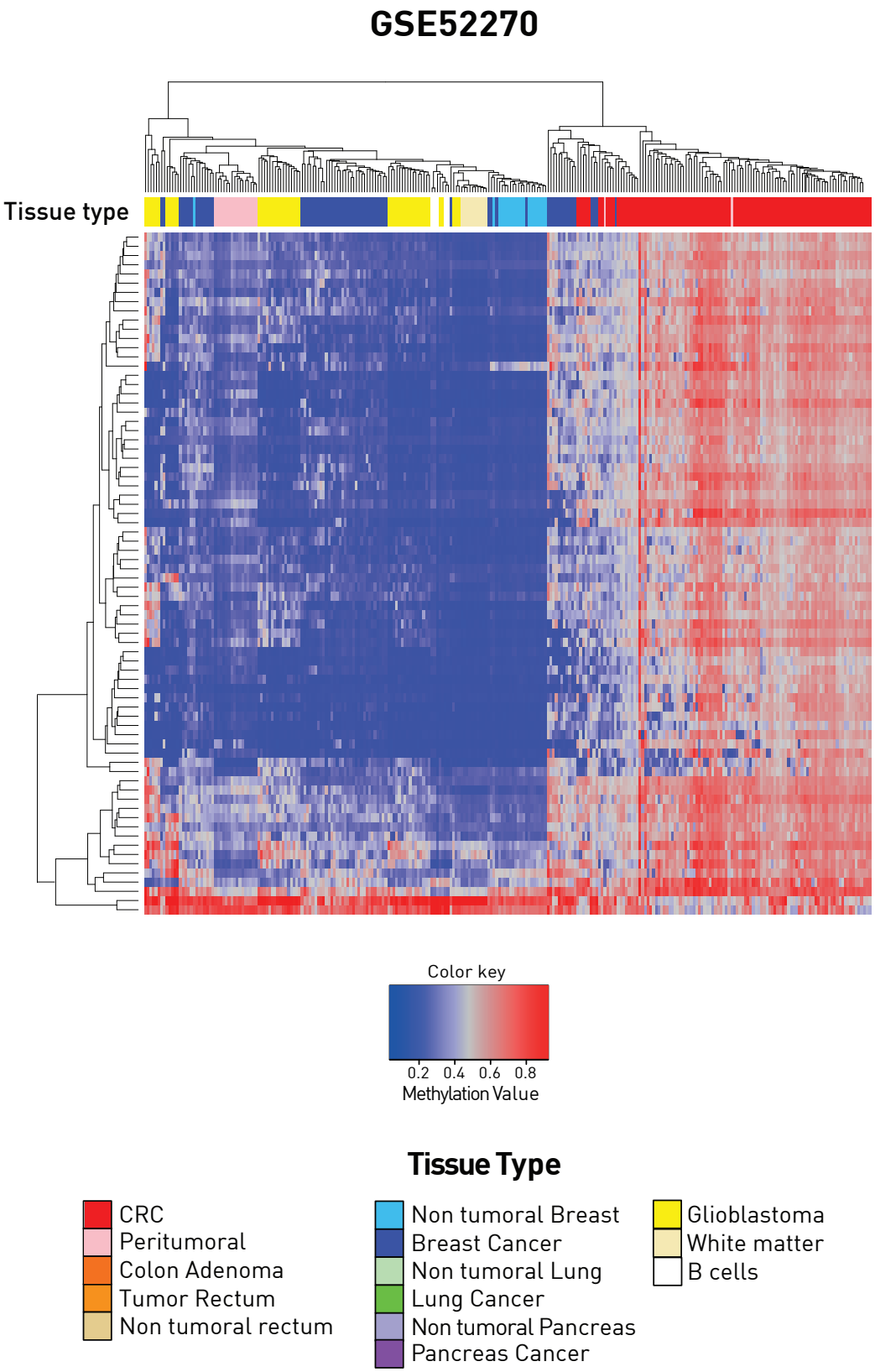

Figure 7: Heat map obtained by an unsupervised hierarchical clustering analysis performed on the GSE52270 data set by using the average methylation beta value for each of the 74 CpG islands.

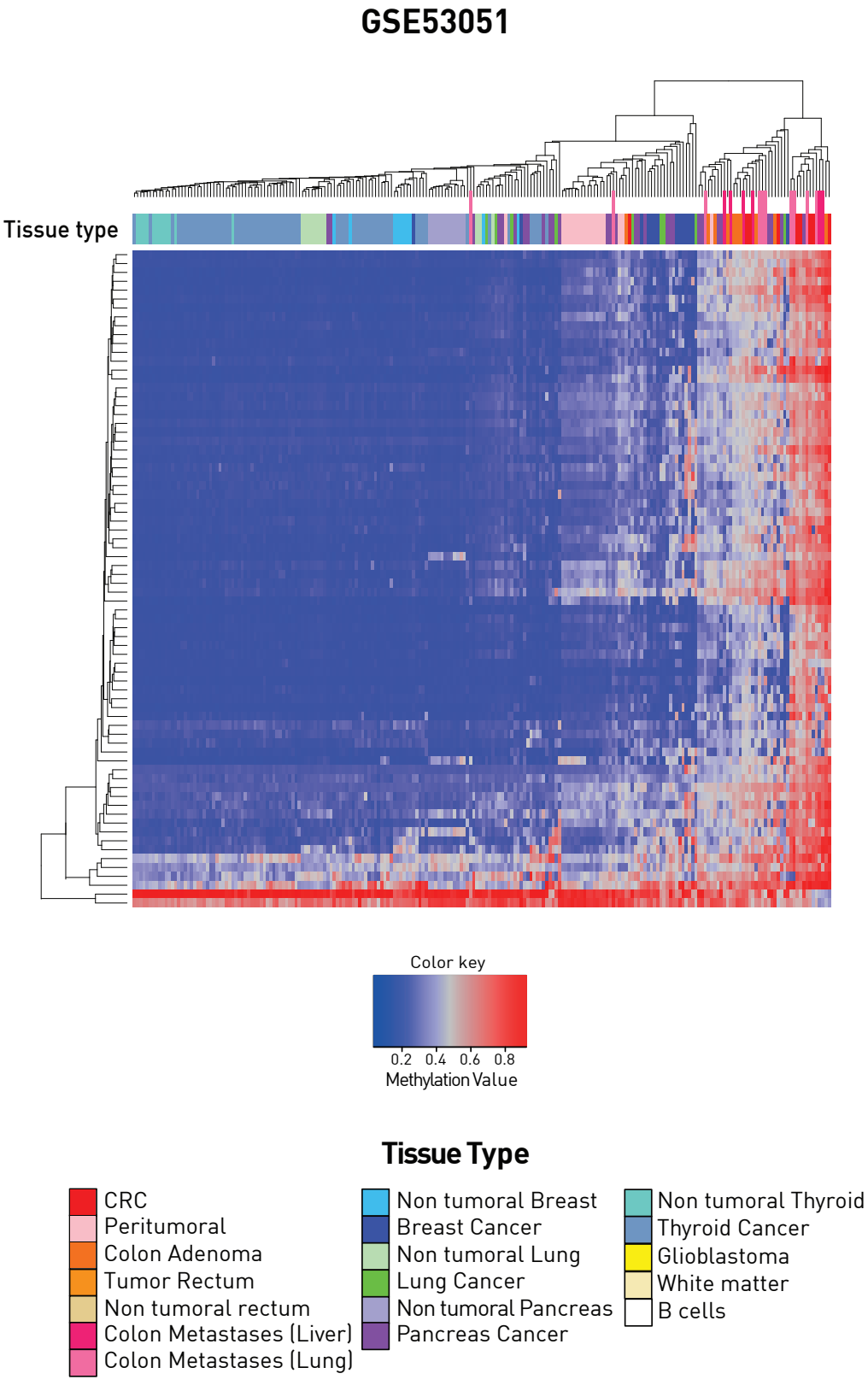

Figure 8: Heat map obtained by an unsupervised hierarchical clustering analysis performed on the GSE53051 data set by using the average methylation beta value for each of the 74 CpG islands.

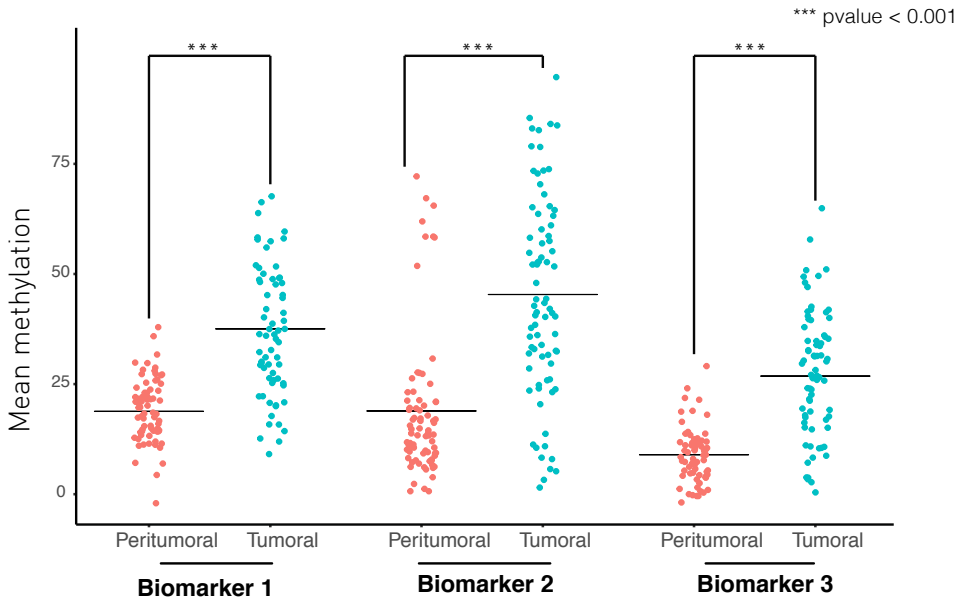

Figure 9: Pyrosequencing methylation analysis of three selected islands, performed in a second data set of 78 tumoral and 78 peritumoral samples.

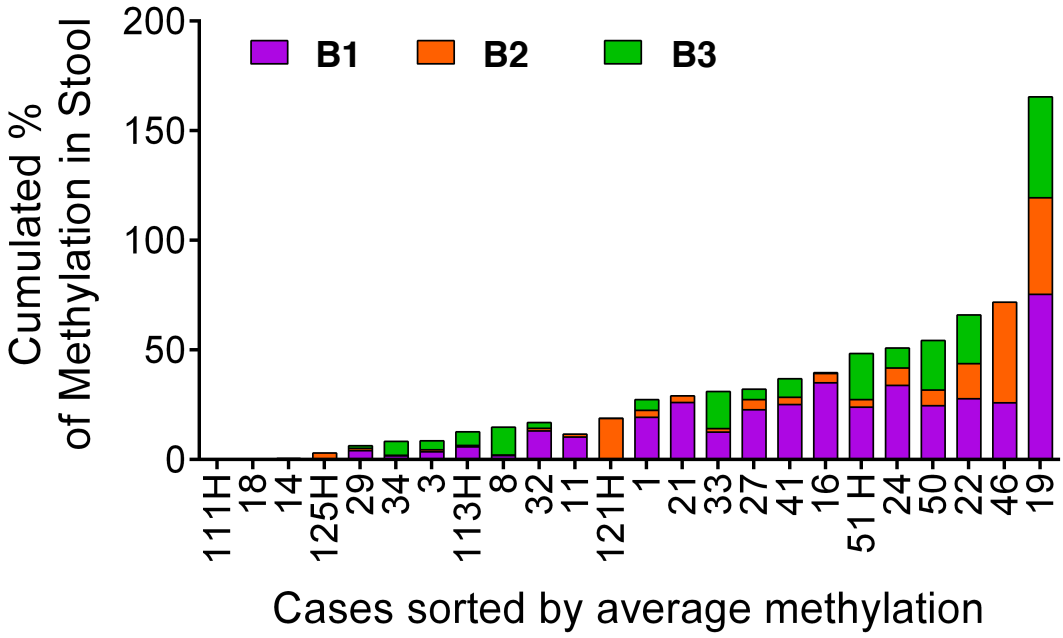

Figure 10: MethylBEAMing analysis results obtained for the three selected biomarkers in DNA isolated from stool samples of CRC patients, taken at the time of surgical resection. Colored bars show the methylation percentage at the three islands, cumulated for each sample.

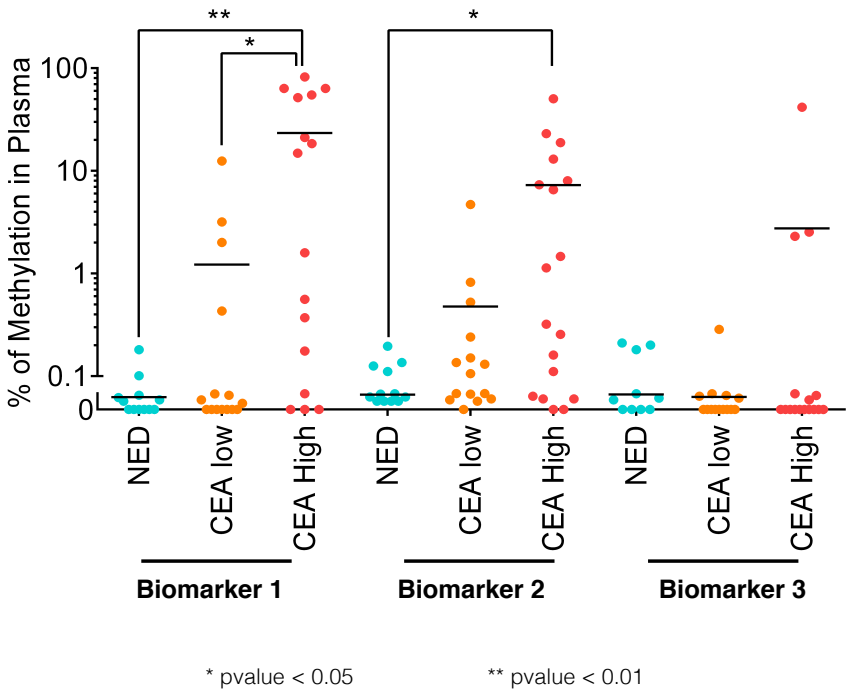

Figure 11: MethylBEAMing methylation value in cfDNA, isolated from plasma, for the three selected biomarkers. Samples were divided into three groups: NED, CEA-low and CEA-high, using a threshold of 5 ng/dl.

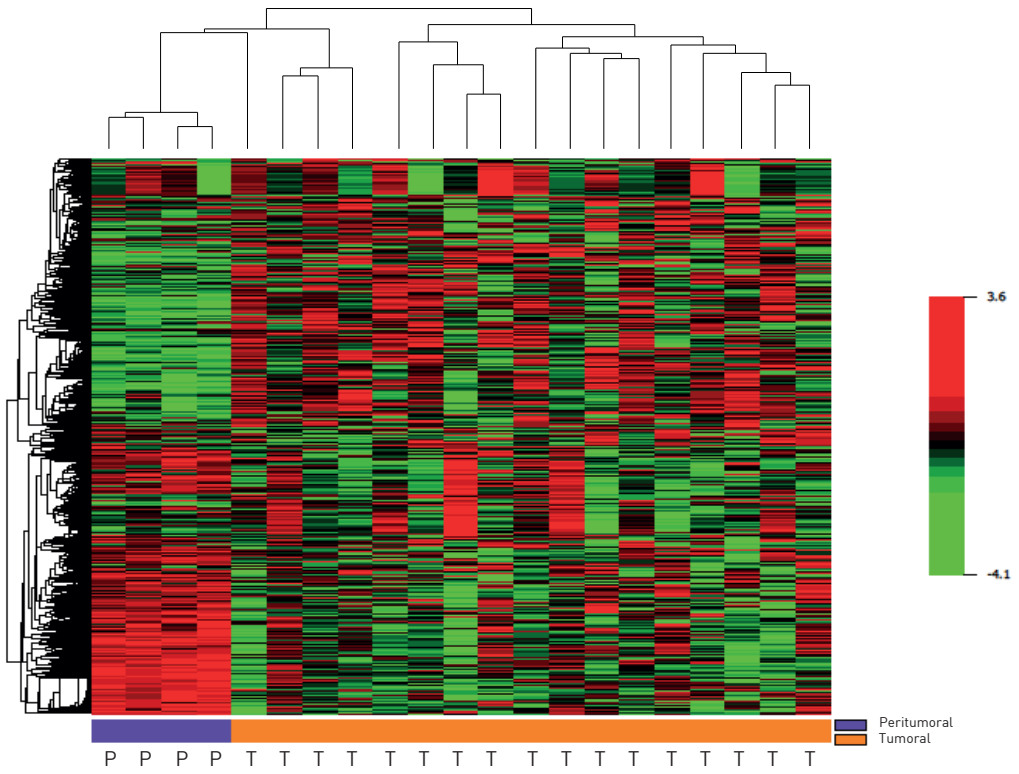

Figure 12: Heat map showing the results of an unsupervised hierarchical clustering analysis performed on the average intensity values obtained from the WGGE analysis on the discovery CRC and peritumoral tissue samples set

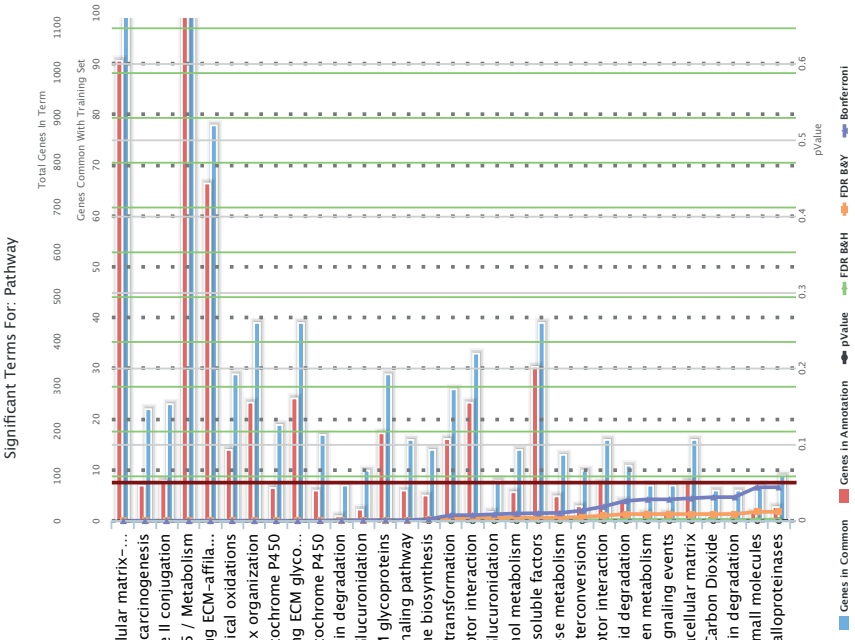

Figure 13: Pathways resulting disregulated by functional annotation of genes differentially expressed in the WGGE analysis, performed by ToppGene package.

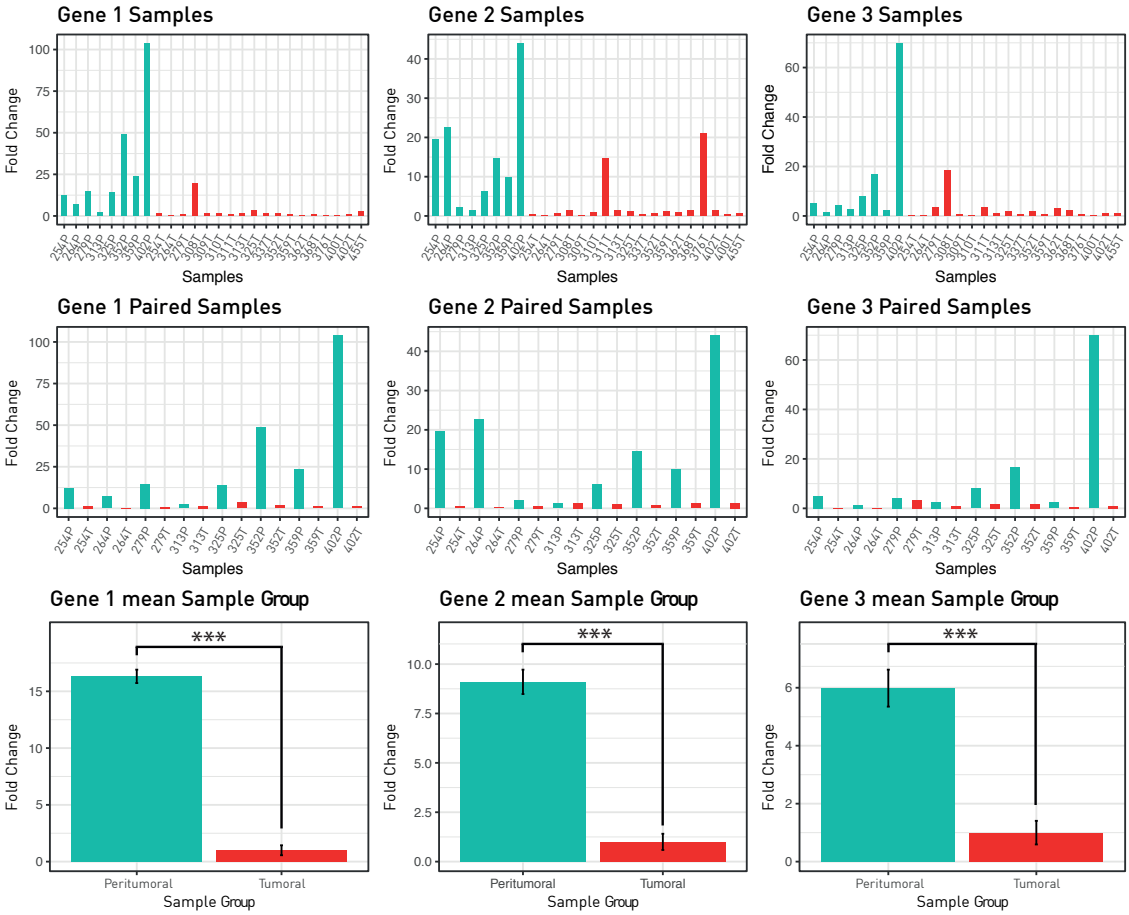

Figure 14: Validation of changes in gene expression observed by WGGE analysis, evaluated by qRT-PCR on three selected genes

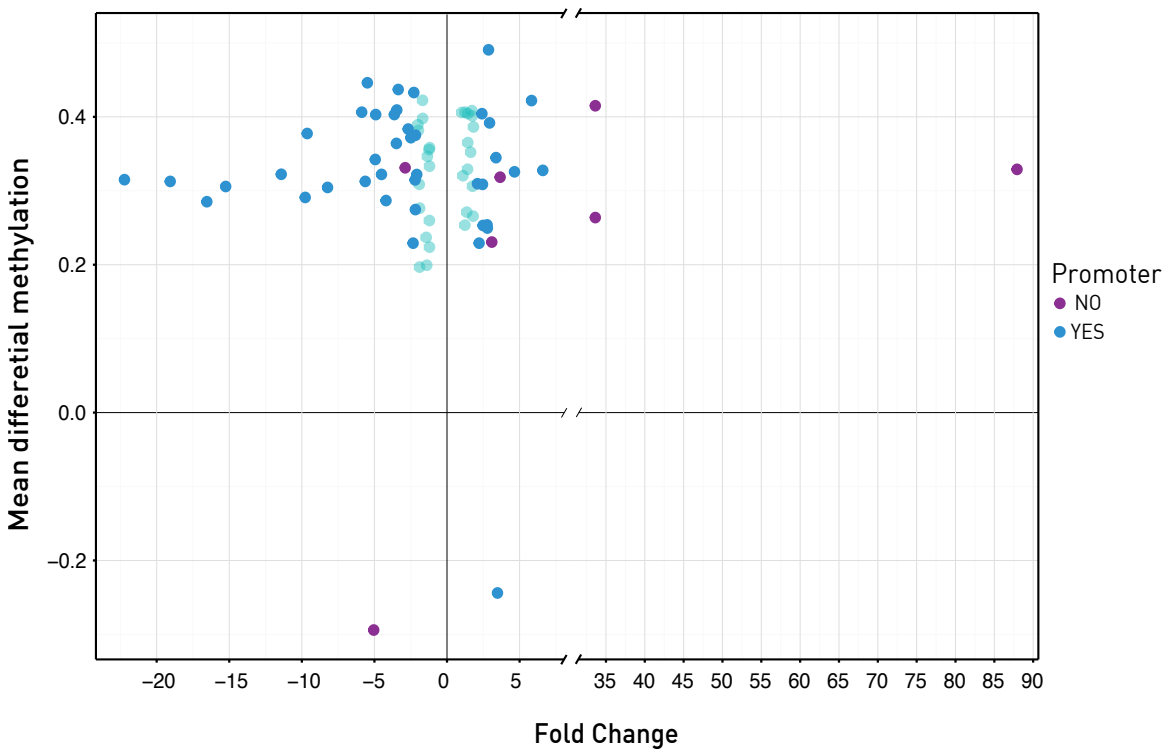

Figure 15: TCGA CRC RNAseq data corresponding to the genes neighbors the 74 selected altered CpG islands. The X axis shows the fold change value, along the Y axis the average value of differential methylation for each CpG island. In blue CpG islands included in promoter regions (according to the definition given in Methods), in purple those not falling into promoters. More faded the symbols corresponding to the CpG islands showing a not significant fold change (between -2 and +2).

Name	Source	p-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y	Hit Count in Query List	Hit Count in Genome
Wnt signaling pathway	PantherDB	1,666E-10	2,128E-07	1,459E-07	0,000001128	35	305
Neuronal System	BioSystems: REACTOME	2,284E-10	2,917E-07	1,459E-07	0,000001128	34	293
Cadherin signaling pathway	PantherDB	5,737E-10	7,327E-07	2,442E-07	0,000001888	24	159
Transmission across Chemical Synapses	BioSystems: REACTOME	2,968E-09	0,00000379	9,476E-07	0,000007325	26	200
Neuroactive ligand-receptor interaction	BioSystems: KEGG	4,458E-08	0,00005693	0,00001139	0,00008801	29	275
Neurotransmitter Release Cycle	BioSystems: REACTOME	2,528E-07	0,0003228	0,00005379	0,0004158	10	37
GABAergic synapse	BioSystems: KEGG	0,000001633	0,002085	0,0002978	0,002302	14	90
Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	MSigDB C2 BIOCARTA (v5.1)	0,000001871	0,002389	0,0002986	0,002308	26	275
Calcium signaling pathway	BioSystems: KEGG	0,000002823	0,003605	0,0004006	0,003097	20	181
Cholinergic synapse	BioSystems: KEGG	0,00000537	0,006858	0,0006858	0,005301	15	113
Metabotropic glutamate receptor group III pathway	PantherDB	0,000008014	0,01023	0,0009304	0,007192	11	64
Ionotropic glutamate receptor pathway	PantherDB	0,00001821	0,02325	0,001938	0,01498	9	46
Na dependent neurotransmitter transporters	BioSystems: REACTOME	0,00002594	0,03312	0,002238	0,0173	6	19
GABA synthesis, release, reuptake and degradation	BioSystems: REACTOME	0,00002594	0,03312	0,002238	0,0173	6	19
Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	MSigDB C2 BIOCARTA (v5.1)	0,00002748	0,03509	0,002238	0,0173	59	1028
Extracellular matrix organization	BioSystems: REACTOME	0,00002804	0,03581	0,002238	0,0173	23	264
Glutamatergic synapse	BioSystems: KEGG	0,00003312	0,04229	0,002398	0,01854	14	116
Cell adhesion molecules (CAMs)	BioSystems: KEGG	0,0000338	0,04316	0,002398	0,01854	16	147

Table 3: Pathways resulting disregulated by functional annotation of genes differentially methylated in the methylome analysis in CRC samples vs peritumoral samples, performed by ToppGene package.

Name	Source	p-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y	Hit Count in Query List	Hit Count in Genome
Cadherin signaling pathway	PantherDB	3,96E-26	7,884E-23	7,884E-23	6,445E-22	60	159
Wnt signaling pathway	PantherDB	1,436E-20	2,858E-17	1,429E-17	1,168E-16	77	305
Neuroactive ligand-receptor interaction	BioSystems: KEGG	3,734E-17	7,434E-14	2,478E-14	2,025E-13	67	275
Neuronal System	BioSystems: REACTOME	1,944E-12	3,871E-09	9,678E-10	7,91E-09	61	293
Neural Crest Differentiation	BioSystems: WikiPathways	6,197E-10	0,000001234	2,467E-07	0,000002017	29	101
Extracellular matrix organization	BioSystems: REACTOME	2,137E-09	0,000004255	7,092E-07	0,000005797	51	264
Transmission across Chemical Synapses	BioSystems: REACTOME	4,476E-09	0,000008912	0,000001273	0,00001041	42	200
Calcium signaling pathway	BioSystems: KEGG	2,466E-08	0,0000491	0,000006138	0,00005017	38	181
GPCR ligand binding	BioSystems: REACTOME	5,407E-07	0,001076	0,0001196	0,0009777	66	445
Nicotine addiction	BioSystems: KEGG	0,000001259	0,002506	0,0002506	0,002048	14	40
Cardiac Progenitor Differentiation	BioSystems: WikiPathways	0,000002122	0,004224	0,000384	0,003139	16	53
G protein signaling via Galphaq family	Pathway Ontology	0,000003122	0,006217	0,0004818	0,003938	8	14
G alpha (q) signalling events	BioSystems: REACTOME	0,000003355	0,00668	0,0004818	0,003938	35	193
Voltage gated Potassium channels	BioSystems: REACTOME	0,000003388	0,006745	0,0004818	0,003938	14	43
Glutamatergic synapse	BioSystems: KEGG	0,000003698	0,007362	0,0004908	0,004012	25	116
Class A/1 (Rhodopsin-like receptors)	BioSystems: REACTOME	0,000004646	0,009251	0,0005782	0,004726	49	316
Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	MSigDB C2 BIOCARTA (v5.1)	0,00000627	0,01248	0,0007343	0,006002	44	275
Metabotropic glutamate receptor group III pathway	PantherDB	0,000007156	0,01425	0,0007915	0,00647	17	64
Morphine addiction	BioSystems: KEGG	0,00001022	0,02034	0,001071	0,008752	21	93
Maturity onset diabetes of the young	BioSystems: KEGG	0,00001083	0,02156	0,001078	0,008811	10	25
Developmental Biology	BioSystems: REACTOME	0,00001148	0,02287	0,001089	0,0089	59	419
Peptide ligand-binding receptors	BioSystems: REACTOME	0,00001456	0,02899	0,001318	0,01077	33	189
Neurotransmitter Release Cycle	BioSystems: REACTOME	0,0000178	0,03544	0,001541	0,01259	12	37
Circadian entrainment	BioSystems: KEGG	0,0000203	0,04043	0,001684	0,01377	21	97
GABAergic synapse	BioSystems: KEGG	0,00002123	0,04227	0,001691	0,01382	20	90
Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	MSigDB C2 BIOCARTA (v5.1)	0,00002284	0,04547	0,001749	0,0143	118	1028
Cholinergic synapse	BioSystems: KEGG	0,00002405	0,04788	0,001773	0,0145	23	113

Table 4: Pathways resulting disregulated by functional annotation of genes differentially methylated in the methylome analysis in adenoma samples vs normal samples mucosa, performed by ToppGene package.

	Adenomas	Controls	Adenomas Δβ value	CRC	Peritumoral	CRC Δβ value	CRC-Adenomas ΔΔβ value
Biomarker 1	0,55	0,35	0,20	0,67	0,39	0,29	0,08
Biomarker 2	0,51	0,29	0,22	0,59	0,35	0,25	0,03
Biomarker 3	0,49	0,27	0,21	0,58	0,29	0,29	0,08
Biomarker 4	0,50	0,34	0,16	0,58	0,39	0,20	0,04
Biomarker 5	0,40	0,16	0,24	0,53	0,17	0,36	0,12
Biomarker 6	0,47	0,21	0,26	0,55	0,24	0,31	0,05
Biomarker 7	0,50	0,28	0,22	0,61	0,30	0,31	0,09
Biomarker 8	0,53	0,17	0,36	0,63	0,18	0,45	0,08
Biomarker 9	0,41	0,18	0,23	0,56	0,19	0,37	0,14
Biomarker 10	0,49	0,24	0,25	0,60	0,24	0,36	0,11
Biomarker 11	0,50	0,33	0,18	0,61	0,18	0,43	0,25
Biomarker 12	0,40	0,19	0,21	0,50	0,19	0,30	0,09
Biomarker 13	0,46	0,25	0,21	0,55	0,23	0,32	0,11
Biomarker 14	0,45	0,19	0,26	0,59	0,19	0,40	0,14
Biomarker 15	0,53	0,27	0,26	0,59	0,36	0,23	-0,03
Biomarker 16	0,35	0,11	0,24	0,49	0,08	0,41	0,16
Biomarker 17	0,25	0,06	0,19	0,38	0,05	0,32	0,13
Biomarker 18	0,49	0,23	0,26	0,58	0,24	0,34	0,08
Biomarker 19	0,64	0,83	-0,19	0,63	0,87	-0,24	-0,05
Biomarker 20	0,48	0,24	0,24	0,58	0,21	0,38	0,13
Biomarker 21	0,52	0,32	0,20	0,63	0,32	0,31	0,11
Biomarker 22	0,65	0,85	-0,20	0,60	0,89	-0,29	-0,09
Biomarker 23	0,42	0,17	0,25	0,58	0,17	0,42	0,16
Biomarker 24	0,63	0,39	0,23	0,73	0,46	0,26	0,03
Biomarker 25	0,53	0,34	0,19	0,63	0,33	0,31	0,11
Biomarker 26	0,43	0,23	0,20	0,52	0,26	0,27	0,06
Biomarker 27	0,44	0,25	0,18	0,56	0,25	0,31	0,12
Biomarker 28	0,50	0,26	0,24	0,64	0,24	0,40	0,17
Biomarker 29	0,48	0,16	0,32	0,61	0,19	0,42	0,10
Biomarker 30	0,56	0,25	0,30	0,69	0,31	0,38	0,08
Biomarker 31	0,56	0,37	0,19	0,66	0,39	0,28	0,09
Biomarker 32	0,48	0,28	0,19	0,57	0,30	0,27	0,08
Biomarker 33	0,59	0,43	0,16	0,69	0,46	0,23	0,07
Biomarker 34	0,58	0,41	0,17	0,69	0,44	0,25	0,08
Biomarker 35	0,48	0,17	0,30	0,64	0,15	0,49	0,19
Biomarker 36	0,44	0,18	0,26	0,57	0,16	0,41	0,14
Biomarker 37	0,49	0,25	0,24	0,55	0,21	0,33	0,09

Biomarker 38	0,40	0,19	0,22	0,51	0,17	0,34	0,13
Biomarker 39	0,55	0,39	0,16	0,65	0,42	0,22	0,06
Biomarker 40	0,49	0,31	0,18	0,57	0,31	0,25	0,07
Biomarker 41	0,43	0,26	0,17	0,56	0,30	0,26	0,09
Biomarker 42	0,62	0,38	0,24	0,70	0,47	0,24	0,00
Biomarker 43	0,45	0,15	0,30	0,54	0,19	0,35	0,05
Biomarker 44	0,41	0,19	0,22	0,54	0,16	0,38	0,16
Biomarker 45	0,38	0,16	0,22	0,47	0,16	0,32	0,09
Biomarker 46	0,42	0,20	0,22	0,56	0,16	0,40	0,18
Biomarker 47	0,47	0,28	0,19	0,59	0,28	0,32	0,13
Biomarker 48	0,45	0,18	0,27	0,59	0,18	0,41	0,14
Biomarker 49	0,39	0,17	0,23	0,53	0,17	0,37	0,14
Biomarker 50	0,64	0,38	0,26	0,74	0,44	0,31	0,05
Biomarker 51	0,50	0,23	0,28	0,65	0,22	0,43	0,16
Biomarker 52	0,43	0,23	0,20	0,58	0,25	0,33	0,13
Biomarker 53	0,33	0,14	0,18	0,48	0,15	0,33	0,15
Biomarker 54	0,49	0,19	0,30	0,60	0,19	0,40	0,10
Biomarker 55	0,34	0,13	0,20	0,51	0,12	0,39	0,19
Biomarker 56	0,62	0,46	0,16	0,70	0,50	0,20	0,04
Biomarker 57	0,35	0,12	0,22	0,52	0,10	0,42	0,20
Biomarker 58	0,36	0,11	0,25	0,49	0,10	0,39	0,15
Biomarker 59	0,48	0,25	0,23	0,59	0,28	0,31	0,08
Biomarker 60	0,43	0,16	0,27	0,52	0,13	0,39	0,12
Biomarker 61	0,44	0,15	0,29	0,57	0,14	0,44	0,15
Biomarker 62	0,46	0,22	0,25	0,59	0,27	0,32	0,08
Biomarker 63	0,49	0,26	0,23	0,59	0,26	0,33	0,10
Biomarker 64	0,48	0,23	0,26	0,64	0,23	0,41	0,15
Biomarker 65	0,35	0,13	0,22	0,48	0,15	0,33	0,11
Biomarker 66	0,37	0,15	0,23	0,49	0,13	0,36	0,13
Biomarker 67	0,51	0,17	0,34	0,67	0,27	0,40	0,06
Biomarker 68	0,44	0,18	0,26	0,56	0,18	0,38	0,12
Biomarker 69	0,49	0,29	0,20	0,60	0,35	0,25	0,05
Biomarker 70	0,52	0,28	0,24	0,62	0,34	0,29	0,04
Biomarker 71	0,45	0,22	0,23	0,57	0,25	0,33	0,09
Biomarker 72	0,52	0,25	0,27	0,59	0,27	0,32	0,05
Biomarker 73	0,38	0,20	0,17	0,51	0,24	0,27	0,10
Biomarker 74	0,33	0,08	0,25	0,43	0,08	0,35	0,10

Table 5: List of biomarkers whose CpG islands were altered both in CRC and in adenomas; for each CpG island are shown the methylation and differential methylation values.

	AUC	Sensibility	Specificity		AUC	Sensibility	Specificity
Biomarker 1	0,98	0,91	0,98	**Biomarker 38**	0,92	0,80	1,00
Biomarker 2	0,96	0,86	1,00	**Biomarker 39**	0,97	0,88	0,98
Biomarker 3	0,97	0,94	0,98	**Biomarker 40**	0,97	0,92	1,00
Biomarker 4	0,98	0,96	1,00	**Biomarker 41**	0,97	0,88	0,98
Biomarker 5	0,98	0,93	1,00	**Biomarker 42**	0,96	0,79	1,00
Biomarker 6	0,99	0,95	1,00	**Biomarker 43**	0,97	0,91	1,00
Biomarker 7	0,94	0,87	0,91	**Biomarker 44**	0,96	0,93	1,00
Biomarker 8	0,97	0,92	0,98	**Biomarker 45**	0,97	0,91	0,95
Biomarker 9	0,97	0,90	1,00	**Biomarker 46**	0,97	0,96	1,00
Biomarker 10	0,99	0,97	0,98	**Biomarker 47**	0,98	0,96	1,00
Biomarker 11	0,92	0,79	0,91	**Biomarker 48**	0,98	0,94	1,00
Biomarker 12	0,97	0,93	1,00	**Biomarker 49**	0,98	0,95	1,00
Biomarker 13	0,96	0,90	0,98	**Biomarker 50**	0,98	0,96	0,98
Biomarker 14	0,98	0,94	0,95	**Biomarker 51**	1,00	0,96	1,00
Biomarker 15	0,90	0,85	0,89	**Biomarker 52**	0,98	0,92	1,00
Biomarker 16	0,95	0,91	1,00	**Biomarker 53**	0,97	0,94	0,95
Biomarker 17	0,96	0,89	1,00	**Biomarker 54**	0,96	0,92	1,00
Biomarker 18	0,97	0,93	0,98	**Biomarker 55**	0,96	0,89	1,00
Biomarker 19	0,98	0,97	1,00	**Biomarker 56**	0,93	0,70	1,00
Biomarker 20	1,00	0,97	0,95	**Biomarker 57**	0,96	0,91	0,98
Biomarker 21	0,99	0,96	1,00	**Biomarker 58**	0,93	0,86	1,00
Biomarker 22	0,94	0,86	0,95	**Biomarker 59**	0,95	0,89	0,98
Biomarker 23	0,98	0,94	0,98	**Biomarker 60**	0,96	0,90	1,00
Biomarker 24	0,93	0,81	0,95	**Biomarker 61**	0,98	0,94	1,00
Biomarker 25	0,97	0,90	1,00	**Biomarker 62**	0,96	0,88	1,00
Biomarker 26	0,93	0,82	0,95	**Biomarker 63**	0,98	0,96	1,00
Biomarker 27	0,98	0,92	1,00	**Biomarker 64**	0,99	0,97	1,00
Biomarker 28	0,91	0,78	0,95	**Biomarker 65**	0,96	0,91	1,00
Biomarker 29	0,93	0,75	0,98	**Biomarker 66**	0,96	0,91	1,00
Biomarker 30	0,99	0,96	0,98	**Biomarker 67**	0,99	0,96	1,00
Biomarker 31	0,89	0,77	0,98	**Biomarker 68**	0,99	0,95	0,98
Biomarker 32	0,99	0,95	0,98	**Biomarker 69**	0,98	0,89	0,98
Biomarker 33	0,97	0,91	1,00	**Biomarker 70**	0,98	0,94	1,00
Biomarker 34	0,94	0,86	0,95	**Biomarker 71**	0,99	0,97	1,00
Biomarker 35	0,96	0,93	1,00	**Biomarker 72**	0,96	0,92	0,95
Biomarker 36	0,99	0,96	0,98	**Biomarker 73**	0,96	0,91	0,95
Biomarker 37	0,97	0,92	0,95	**Biomarker 74**	0,96	0,94	0,98

Table 6: List of AUC, specificity and sensitivity for each biomarker calculated in the TCGA dataset

Bibliography

1. Holliday, R. The inheritance of epigenetic defects. Science 238, 163–70 (1987).
2. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol 20, 274–281 (2013).
3. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–92 (2012).
4. McCabe, M. T., Lee, E. K. & Vertino, P. M. A multifactorial signature of DNA sequence and polycomb binding predicts aberrant CpG island methylation. Cancer Res. 69, 282–291 (2009).
5. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. J. Mol. Biol. 196, 261–282 (1987).
6. Luger, K., Mäder, a W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389, 251–260 (1997).
7. Kouzarides, T. Chromatin Modifications and Their Function. Cell 128, 693–705 (2007).
8. Jenuwein, T. & Allis, C. D. Translating the histone code. Science (80-.). 293, 1074–1080 (2001).
9. Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. EMBO J. 7, 1395–1402 (1988).
10. Liang, G. et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc. Natl. Acad. Sci. U. S. A. 101, 7357–7362 (2004).
11. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560 (2007).
12. Bernstein, B. E. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125, 315–326 (2006).
13. Ringrose, L. & Paro, R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development 134, 223–32 (2007).
14. Haberland, M., Montgomery, R. L. & Olson, E. N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. Nat. Rev. Genet. 10, 32–42 (2009).
15. Shi, Y. Histone lysine demethylases: emerging roles in development, physiology and disease. Nat. Rev. Genet. 8, 829–33 (2007).
16. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. 10, 295–304 (2009).
17. Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H. & Shinkai, Y. G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. EMBO J. 27, 2681–90 (2008).
18. 1Lehnertz, B. et al. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr. Biol. 13, 1192–1200 (2003).
19. Zhao, Q. et al. PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. Nat. Struct. Mol. Biol. 16, 304–311 (2009).
20. Estève, P.-O. et al. Regulation of DNMT1 stability through SET7-mediated lysine methylation in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. 106, 5076–81 (2009).
21. Wang, J. et al. The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. Nat. Genet. 41, 125–9 (2009).
22. Fuks, F. et al. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J. Biol. Chem. 278, 4035–4040 (2003).
23. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. Nature 403, 501–2 (2000).
24. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science 334, 194 (2011).
25. Kochanek, S., Renz, D. & Doerfler, W. Transcriptional silencing of human Alu sequences and inhibition of protein binding in the box B regulatory elements by 5???-CG-3??? methylation. FEBS Lett. 360, 115–120 (1995).

26. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics 13, 335–340 (1997).
27. Robertson, K. D. DNA methylation and human disease. Nat. Rev. Genet. 6, 597–610 (2005).
28. Jelinic, P. & Shaw, P. Loss of imprinting and cancer. Journal of Pathology 211, 261–268 (2007).
29. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247–257 (1999).
30. Straussman, R. et al. Developmental programming of CpG island methylation profiles in the human genome. Nat. Struct. Mol. Biol. 16, 564–571 (2009).
31. Laurent, L. et al. Dynamic changes in the human methylome during differentiation. Genome Res. 20, 320–331 (2010).
32. Brandeis, M. et al. Sp1 elements protect a CpG island from de novo methylation. Nature 371, 435–438 (1994).
33. Weber, M. et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. 39, 457–66 (2007).
34. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454, 766–770 (2008).
35. Ooi, S. K. T. et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 448, 714–717 (2007).
36. Otani, J. et al. Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. EMBO Rep 10, 1235–1241 (2009).
37. Pollack, Y., Stein, R., Razin, A. & Cedar, H. Methylation of foreign DNA sequences in eukaryotic cells. Proc. Natl. Acad. Sci. U. S. A. 77, 6463–7 (1980).
38. Leonhardt, H., Page, a W., Weier, H. U. & Bestor, T. H. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell 71, 865–873 (1992).
39. Gruenbaum, Y., Cedar, H. & Razin, A. Substrate and sequence specificity of a eukaryotic DNA methylase. Nature 295, 620–622 (1982).
40. Ben-Shushan, E., Pikarsky, E., Klar, A. & Bergman, Y. Extinction of Oct-3/4 gene expression in embryonal carcinoma x fibroblast somatic cell hybrids is accompanied by changes in the methylation status, chromatin structure, and transcriptional activity of the Oct-3/4 upstream region. Mol. Cell. Biol. 13, 891–901 (1993).
41. Gidekel, S. & Bergman, Y. A unique developmental pattern of Oct-3/4 DNA methylation is controlled by a cis-demodification element. J. Biol. Chem. 277, 34521–34530 (2002).
42. Feldman, N. et al. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat. Cell Biol. 8, 188–194 (2006).
43. Epsztejn-Litman, S. et al. De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat. Struct. Mol. Biol. 15, 1176–83 (2008).
44. Keohane, A. M., Lavender, J. S., O'Neill, L. P. & Turner, B. M. Histone acetylation and X inactivation. Dev Genet 22, 65–73 (1998).
45. Plath, K. et al. Role of histone H3 lysine 27 methylation in X inactivation. Science 300, 131–5 (2003).
46. Silva, J. et al. Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. Developmental Cell 4, 481–495 (2003).
47. Kaslow, D. C. & Migeon, B. R. DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. Proc. Natl. Acad. Sci. U. S. A. 84, 6210–4 (1987).
48. Wareham, K. A., Lyon, M. F., Glenister, P. H. & Williams, E. D. Age related reactivation of an X-linked gene. Nature 327, 725–727 (1987).
49. Illingworth, R. S. & Bird, A. P. CpG islands - 'A rough guide'. FEBS Letters 583, 1713–1720 (2009).
50. Viré, E. et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439, 871–874 (2006).
51. Hitchins, M. P. et al. Dominantly Inherited Constitutional Epigenetic Silencing of MLH1 in a Cancer-Affected Family Is Linked to a Single Nucleotide Variant within the 5'UTR. Cancer Cell 20, 200–213 (2011).

52. Boumber, Y. A. et al. An Sp1/Sp3 binding polymorphism confers methylation protection. PLoS Genet. 4, (2008).
53. Wu, H. et al. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473, 389–393 (2011).
54. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat. Commun. 2, 241 (2011).
55. Yisraeli, J. et al. Muscle-specific activation of a methylated chimeric actin gene. Cell 46, 409–416 (1986).
56. Busslinger, M., Hurst, J. & Flavell, R. A. DNA methylation and the regulation of globin gene expression. Cell 34, 197–206 (1983).
57. Goren, A. et al. Fine tuning of globin gene expression by DNA methylation. PLoS One 1, (2006).
58. Siegfried, Z. & Cedar, H. DNA methylation: a molecular lock. Curr. Biol. 7, R305–R307 (1997).
59. Qian, W. et al. A Histone Acetyltransferase Regulates Active DNA Demethylation in Arabidopsis. Science (80-.). 336, 1445–1448 (2012).
60. Venter, J. C. The Sequence of the Human Genome. Science (80-.). 291, 1304–1351 (2001).
61. Kelly, T. K. et al. H2A.Z maintenance during mitosis reveals nucleosome shifting on mitotically silenced genes. Mol. Cell 39, 901–911 (2010).
62. Taberlay, P. C. et al. Polycomb-repressed genes have permissive enhancers that initiate reprogramming. Cell 147, 1283–1294 (2011).
63. Shi, Y. et al. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119, 941–953 (2004).
64. Takeuchi, T., Watanabe, Y., Takano-Shimizu, T. & Kondo, S. Roles of jumonji and jumonji family genes in chromatin regulation and development. Developmental Dynamics 235, 2449–2459 (2006).
65. Comb, M. & Goodman, H. M. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. 18, 3975–3982 (1990).
66. Prendergast, G. C., Lawe, D. & Ziff, E. B. Association of Myn, the murine homolog of Max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. Cell 65, 395–407 (1991).
67. Jones, P. L. et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet. 19, 187–191 (1998).
68. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. Cell 128, 707–719 (2007).
69. Illingworth, R. S. et al. Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. PLoS Genet. 6, (2010).
70. Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T. & Migeon, B. R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: Implications for X chromosome inactivation. Proc. Natl. Acad. Sci. U. S. A. 81, (1984).
71. Larsen, F., Solheim, J. & Prydz, H. A methylated CpG Island 3' in the apolipoprotein-E gene does not repress its transcription. Hum. Mol. Genet. 2, 775–780 (1993).
72. Maunakea, A. K. et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466, 253–7 (2010).
73. Nguyen, C. et al. Susceptibility of nonpromoter CpG islands to de novo methylation in normal and neoplastic cells. J. Natl. Cancer Inst. 93, 1465–1472 (2001).
74. Nguyen, C. T., Gonzales, F. A. & Jones, P. A. Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. Nucleic Acids Res. 29, 4598–606 (2001).
75. Jones, P. A. The DNA methylation paradox. Trends in Genetics 15, 34–37 (1999).
76. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–5 (2011).
77. Recillas-Targa, F. et al. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. Proc. Natl. Acad. Sci. U. S. A. 99, 6883–6888 (2002).
78. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405, 482–485 (2000).
79. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. Cell 87, 159–170 (1996).

80. Feinberg, A., Vogelstein, B., Droller, M., Baylin, S. & Nelkin, B. Mutation affecting the 12th amino acid of the c-Ha-ras oncogene product occurs infrequently in human cancer. Science (80-.). 220, 1175–1177 (1983).
81. Wang, Z. et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 304, 1164–6 (2004).
82. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. Cell 128, 683–692 (2007).
83. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301, 89–92 (1983).
84. Jones, P. a & Baylin, S. B. The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. 3, 415–28 (2002).
85. Rodriguez, J. et al. Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res. 66, 8462–9468 (2006).
86. Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. Science (80-.). 300, 2003 (2003).
87. Howard, G., Eiges, R., Gaudet, F., Jaenisch, R. & Eden, a. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene 27, 404–8 (2008).
88. Greger, V., Passarge, E., Höpping, W., Messmer, E. & Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. Hum. Genet. 83, 155–158 (1989).
89. Baylin, S. B. DNA methylation and gene silencing in cancer. Nat. Clin. Pract. Oncol. 2, S4–S11 (2005).
90. Jones, P. a & Laird, P. W. Cancer epigenetics comes of age. Nat. Genet. 21, 163–167 (1999).
91. Long, C. et al. Promoter hypermethylation of the RUNX3 gene in esophageal squamous cell carcinoma. Cancer Invest. 25, 685–690 (2007).
92. Akiyama, Y. et al. GATA-4 and GATA-5 transcription factor genes and potential downstream antitumor target genes are epigenetically silenced in colorectal and gastric cancer. Mol. Cell. Biol. 23, 8429–39 (2003).
93. Berdasco, M. & Esteller, M. Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. Developmental Cell 19, 698–711 (2010).
94. Schlesinger, Y. et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat. Genet. 39, 232–236 (2007).
95. Widschwendter, M. et al. Epigenetic stem cell signature in cancer. Nat. Genet. 39, 157–158 (2007).
96. Ohm, J. E. et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat. Genet. 39, 237–42 (2007).
97. Gal-Yam, E. N. et al. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc. Natl. Acad. Sci. U. S. A. 105, 12979–12984 (2008).
98. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. Carcinogenesis 31, 27–36 (2009).
99. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. U. S. A. 100, 3983–8 (2003).
100. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and Epigenetic Regulators of Pluripotency. Cell 128, 747–762 (2007).
101. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? Nat. Rev. Cancer 6, 107–116 (2006).
102. Cui, H. et al. Loss of imprinting in colorectal cancer linked to hypomethylation of H19 and IGF2. Cancer Res. 62, 6442–6446 (2002).
103. Siegel, R., Desantis, C. & Jemal, A. Colorectal cancer statistics, 2014. CA Cancer J Clin 64, 104–117 (2014).
104. Ferlay, J. et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer 136, E359–E386 (2015).
105. Rustgi, A. K. The genetics of hereditary colon cancer. Genes and Development 21, 2525–2538 (2007).

106. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–767 (1990).
107. Grady, W. M. & Carethers, J. M. Genomic and Epigenetic Instability in Colorectal Cancer Pathogenesis. Gastroenterology 135, 1079–1099 (2008).
108. Herman, J. G. et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. U. S. A. 95, 6870–5 (1998).
109. Weisenberger, D. J. et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. 38, 787–93 (2006).
110. Umar, A. et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J. Natl. Cancer Inst. 96, 261–8 (2004).
111. Giardiello, F. M. et al. Guidelines on genetic evaluation and management of lynch syndrome: A consensus statement by the us multi-society task force on colorectal cancer. Gastroenterology 147, 502–526 (2014).
112. Arends, M. J. Pathways of colorectal carcinogenesis. Appl. Immunohistochem. Mol. Morphol. 21, 97–102 (2013).
113. Gay, L. J. et al. MLH1 promoter methylation, diet, and lifestyle factors in mismatch repair deficient colorectal cancer patients from EPIC-Norfolk. Nutr. Cancer 63, 1000–10 (2011).
114. Issa, J.-P. CpG island methylator phenotype in cancer. Nat. Rev. Cancer 4, 988–993 (2004).
115. Hawkins, N. et al. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. Gastroenterology 122, 1376–87. (2002).
116. van Rijnsoever, M., Grieu, F., Elsaleh, H., Joseph, D. & Iacopetta, B. Characterisation of colorectal cancers showing hypermethylation at multiple CpG islands. Gut 51, 797–802 (2002).
117. Barault, L. et al. Hypermethylator phenotype in sporadic colon cancer: Study on a population-based series of 582 cases. Cancer Res. 68, 8541–8546 (2008).
118. Samowitz, W. S. et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology 129, 837–845 (2005).
119. Nosho, K. et al. Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. PLoS One 3, e3698 (2008).
120. Park, S.-J. et al. Frequent CpG island methylation in serrated adenomas of the colorectum. Am. J. Pathol. 162, 815–822 (2003).
121. Shen, L. et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc. Natl. Acad. Sci. U. S. A. 104, 18654–9 (2007).
122. Suehiro, Y. et al. Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. Clin. Cancer Res. 14, 2560–2569 (2008).
123. Curtin, K., Slattery, M. L. & Samowitz, W. S. CpG island methylation in colorectal cancer: past, present and future. Patholog. Res. Int. 2011, 902674 (2011).
124. Yamashita, K., Dai, T., Dai, Y., Yamamoto, F. & Perucho, M. Genetics supersedes epigenetics in colon cancer phenotype. Cancer Cell 4, 121–131 (2003).
125. Sinicrope, F. A. et al. Prognostic Impact of Microsatellite Instability and DNA Ploidy in Human Colon Carcinoma Patients. Gastroenterology 131, 729–737 (2006).
126. Cheng, Y.-W. et al. CpG island methylator phenotype associates with low-degree chromosomal abnormalities in colorectal cancer. Clin. Cancer Res. 14, 6005–13 (2008).
127. Jass, J. R. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50, 113–130 (2007).
128. Pancione, M., Remo, A. & Colantuoni, V. Genetic and epigenetic events generate multiple pathways in colorectal cancer progression. Patholog. Res. Int. 2012, (2012).
129. Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. 89, 1827–31 (1992).
130. Susan, J. Ci., Harrison, J., Paul, C. L. & Frommer, M. High sensitivity mapping of methylated cytosines. Nucleic Acids Res. 22, 2990–2997 (1994).
131. Weber, M. et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. 37, 853–62 (2005).

132. Serre, D., Lee, B. H. & Ting, A. H. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res. 38, 391–399 (2009).
133. Rauch, T. & Pfeifer, G. P. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. Lab. Invest. 85, 1172–1180 (2005).
134. Nair, S. S. et al. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. Epigenetics 6, 34–44 (2011).
135.1 Robinson, M. D. et al. Evaluation of affinity-based genome-wide DNA methylation data: Effects of CpG density, amplification bias, and copy number variation. Genome Res. 20, 1719–1729 (2010).
136. Gu, H. et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat. Methods 7, 133–6 (2010).
137. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. Genomics 98, 288–295 (2011).
138. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 6, epi.15.114 (2015).
139. The Cancer Genome Network Atlas. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).
140. Eckhardt, F. et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. 38, 1378–1385 (2006).
141. Pidsley, R. et al. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14, 293 (2013).
142. Marabita, F. et al. An evaluation of analysis pipelines for DNA methylation profiling using the illumina humanmethylation450 BeadChip platform. Epigenetics 8, 333–346 (2013).
143. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29, 189–196 (2013).
144. Dedeurwaerder, S. et al. Evaluation of the Infinium Methylation 450K technology. Epigenomics 3, 771–784 (2011).
145. Sun, S., Huang, Y.-W., Yan, P. S., Huang, T. H. & Lin, S. Preprocessing differential methylation hybridization microarray data. BioData Min. 4, 13 (2011).
146. Siegmund, K. D. Statistical approaches for the analysis of DNA methylation microarray data. Human Genetics 129, 585–595 (2011).
147. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13, R44 (2012).
148. Touleimat, N. & Tost, J. Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4, 325–41 (2012).
149. Assenov, Y. et al. Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods 11, 1138–40 (2014).
150. Schalkwyk, L. C., Pidsley, R. & Wong, C. C. Y. wateRmelon: Illumina 450 methylation array normalization and metrics. R package version 1.2.2 (2013).
151. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer 11, 726–34 (2011).
152. Dedeurwaerder, S. et al. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in bioinformatics 15, 929–941 (2014).
153. Fortin, J.-P. et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 15, 503 (2014).
154. Du, P. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11, 587 (2010).
155. Davis S, et al. methylumi: Handle Illumina methylation data. R package version 2.12.0; (2014).
156. Aryee, M. J. et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369 (2014).

157. Morris, T. J. et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics 30, 428–430 (2014).
158. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–27 (2007).
159. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3, 1724–1735 (2007).
160. Naylor, S. Biomarkers : current perspectives. Expert Rev. Mol. diagnostics 3, 525–529 (2003).
161. Mayeux, R. Biomarkers: potential uses and limitations. NeuroRx 1, 182–8 (2004).
162. Strimbu, K. & Tavel, J. a. What are biomarkers? Curr. Opin. HIV AIDS 5, 463–466 (2010).
163. Mandel, J. S. et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. N. Engl. J. Med. 328, 1365–71 (1993).
164. Van Rossum, L. G. M. et al. Colorectal cancer screening comparing no screening, immunochemical and guaiac fecal occult blood tests: A cost-effectiveness analysis. Int. J. Cancer 128, 1908–1917 (2011).
165. Dong, Y., Zhao, H., Li, H., Li, X. & Yang, S. DNA methylation as an early diagnostic marker of cancer (Review). Biomed. Reports 2, 326–330 (2014).
166. Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. Nat. Rev. Gastroenterol. Hepatol. 8, 686–700 (2011).
167. Davies, R. J., Miller, R. & Coleman, N. Colorectal cancer screening: prospects for molecular stool analysis. Nat. Rev. Cancer 5, 199–209 (2005).
168. Ahlquist, D. A., Harrington, J. J., Burgart, L. J. & Roche, P. C. Morphometric analysis of the 'mucocellular layer' overlying colorectal cancer and normal mucosa: Relevance to exfoliation and stool screening. Hum. Pathol. 31, 51–57 (2000).
169. Shirahata, A. et al. Vimentin methylation as a marker for advanced colorectal carcinoma. Anticancer Res. 29, 279–281 (2009).
170. Chen, W. D. et al. Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. J. Natl. Cancer Inst. 97, 1124–1132 (2005).
171. Lofton-Day, C. et al. DNA methylation biomarkers for blood-based colorectal cancer screening. Clin. Chem. 54, 414–423 (2008).
172. J., C. et al. Molecular analysis of APC promoter methylation and protein expression in colorectal cancer metastasis. Carcinogenesis 26, 37–43 (2005).
173. Krasna, M., Flancbaum, L., Cody, R., Shneibaum, S. & Ben Ari, G. Vascular and neural invasion in colorectal carcinoma. Incidence and prognostic significance. Cancer 61, 18–23 (1988).
174. Sastre, J. et al. Circulating tumor cells in colorectal cancer: Correlation with clinical and pathological variables. Ann. Oncol. 19, 935–938 (2008).
175. Wong, I. H. Methylation profiling of human cancers in blood: molecular monitoring and prognostication (review). International journal of oncology 19, 1319–1324 (2001).
176. Johnson, P. J. & Lo, Y. M. Plasma nucleic acids in the diagnosis and management of malignant disease. Clin Chem 48, 1186–1193 (2002).
177. Harris, L. et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J. Clin. Oncol. 25, 5287–312 (2007).
178. Engstrom, P. F. et al. NCCN Colorectal Cancer Practice Guidelines. The National Comprehensive Cancer Network. Oncology (Williston Park). 10, 140–175 (1996).
179. Compton, C., Fenoglio-Preiser, C. M., Pettigrew, N. & Fielding, L. P. American Joint Committee on Cancer Prognostic Factors Consensus Conference: Colorectal Working Group. Cancer 88, 1739–57 (2000).
180. Sidransky, D. Emerging molecular markers of cancer. Nat. Rev. Cancer 2, 210–9 (2002).
181. T., D. et al. Circulating methylated septin 9 DNA in plasma is a biomarker for colorectal cancer. Gastroenterology 136, A623 (2009).
182. Diehl, F. et al. Analysis of Mutations in DNA Isolated From Plasma and Stool of Colorectal Cancer Patients. Gastroenterology 135, (2008).
183. Zou, H. et al. High Detection Rates of Colorectal Neoplasia by Stool DNA Testing With a Novel Digital Melt Curve Assay. Gastroenterology 136, 459–470 (2009).
184. Li, M. et al. Sensitive digital quantification of DNA methylation in clinical samples. Nat. Biotechnol. 27, 858–63 (2009).

185. Diehl, F. et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci U S A 102, 16368–16373 (2005).
186. Olson, J., Whitney, D. H., Durkee, K. & Shuber, A. P. DNA stabilization is critical for maximizing performance of fecal DNA-based colorectal cancer tests. Diagn.Mol.Pathol. 14, 183–191 (2005).
187. Jing, R. R. et al. A sensitive method to quantify human cell-free circulating DNA in blood: Relevance to myocardial infarction screening. Clin. Biochem. 44, 1074–1079 (2011).
188. Barault, L. et al. Digital PCR quantification of MGMT methylation refines prediction of clinical benefit from alkylating agents in glioblastoma and metastatic colorectal cancer. Ann. Oncol. 26, 1994–1999 (2015).
189. Cortes, C. & Vapnik, V. Support-Vector Networks. Mach. Learn. 20, 273–297 (1995).
190. De Benedetti, L. et al. Genetic events in sporadic colorectal adenomas: K-ras and p53 heterozygous mutations are not sufficient for malignant progression. Anticancer Res. 13, 667–670 (1993).
191. De Benedetti, L. et al. Association of APC gene mutations and histological characteristics of colorectal adenomas. Cancer Res 54, 3553–3556 (1994).
192. Lam, K., Pan, K., Linnekamp, J., Medema, J. P. & Kandimalla, R. DNA methylation based biomarkers in colorectal cancer: A systematic review. Biochim. Biophys. Acta 1866, 106–120 (2016).
193. Javier Carmona, F. et al. DNA methylation biomarkers for noninvasive diagnosis of colorectal cancer. Cancer Prev. Res. 6, 656–665 (2013).
194. Luo, Y. et al. Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. Gastroenterology 147, (2014).
195. Heyn, H. et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. 17, 11 (2016).
196. Timp, W. et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med. 6, 61 (2014).