SCIENTIFIC REPORTS
natureresearch

OPEN

DNA sequence context as a marker of CpG methylation instability in normal and cancer tissues

Giovanni Scala[1,5], Antonio Federico[1], Domenico Palumbo[2], Sergio Cocozza[2] & Dario Greco[1,3,4*]

DNA methylation alterations are related to multiple molecular mechanisms. The DNA context of CpG sites plays a crucial role in the maintenance and stability of methylation patterns. The quantitative relationship between DNA composition and DNA methylation has been studied in normal as well as pathological conditions, showing that DNA methylation status is highly dependent on the local sequence context. In this work, we describe this relationship by analyzing the DNA sequence context associated to methylation profiles in both physiological and pathological conditions. In particular, we used DNA motifs to describe methylation stability patterns in normal tissues and aberrant methylation events in cancer lesions. In this manuscript, we show how different groups of DNA sequences can be related to specific epigenetic events, across normal and cancer tissues, and provide a thorough structural and functional characterization of these sequences.

Sizable efforts have characterized epigenetic signatures and their localization across the human genome[1]. However, the mechanisms influencing the sensitivity of certain sites to epigenetic modulation are still not completely understood as many internal and external factors are thought to be involved in this process. This aspect has been extensively studied in bacteria, where particular sequences are able to favor the binding of enzymes that modify the methylation status[2].

It is well known that a dependency exists between DNA-sequence composition and occurrence of methylation in mammalian genomes. A well-known example of DNA-sequence-dependent mechanism is the dependency of DNA methylation on CpGs density[3,4]. Indeed, high density CpG clusters (typically CG Islands - CGIs) are usually located in the proximity of Transcription Start Sites (TSS) and promoters of housekeeping genes in an unmethylated state, hence allowing gene transcription[5,6]. Furthermore, a model of collaborative methylation among CpG sites has been recently suggested, explaining how CpGs in CGIs could maintain a stable methylation through the generations due to their clustered localization[7]. In 2009, McCabe and colleagues identified distinct sequence patterns marking methylation deregulation events in cancer[8].

To date, only a few correlations between phenotypic alterations, DNA methylation alterations and DNA sequence context have been described[9,10]. Ghorbani and colleagues[11] gave an insight to the identification of the motifs underlying common methylation patterns across different cancer types. Their results paved the way to the identification of alternative cancer biomarkers, currently needed in order to expand and diversify the therapeutic possibilities for cancer patients. In this context, epigenetic marks are considered as pivotal elements in biomarker discovery. For example, epigenetic signatures are used as diagnostic, prognostic and therapy-assessment markers in colorectal cancer[12]. However, the genomic context of aberrant methylated CpGs in cancer is still largely unexplored, and it remains unclear whether these contexts can be exploited to track tumor progression.

In this study, we analyzed DNA sequence contexts in the form of DNA motifs characterizing: (i) interindividual CpG methylation variability in normal tissues and (ii) aberrant CpG methylation between normal and cancer tissues. By using such approach we revealed the existence of i) tissue specific motifs characterizing low and high population variability of CpG methylation in normal tissues and (ii) putative cancer-specific biomarkers, constituted by sequence motifs related to aberrant methylation across 33 cancer types. Furthermore, we investigated the regulatory consequences of aberrant methylation in selected genomic contexts, in order to give an insight on the mechanisms underlying the gene expression deregulation in the onset and progression of cancer. Our results

[1]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [2]Department of Molecular Medicine and Medical Biotechnology, University of Naples "Federico II", Naples, Italy. [3]BioMediTech Institute, Tampere University, Tampere, Finland. [4]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [5]Present address: Department of Biology, University of Naples Federico II, 80126, Naples, Italy. *email: dario.greco@tuni.fi

suggest the existence of alternative biomarkers in human cancer and provide novel hints in understanding of the role of the DNA sequence context in genetic and epigenetic deregulation.

Materials and Methods

Data sources.

To investigate the relationship between DNA context and CpG methylation stability, we retrieved pre-processed DNA methylation data by Illumina Infinium 450 k assays from The Cancer Genome Atlas (TCGA)[13] repository and the Epic Italy cohort[14]. Supplementary Table S1 reports, for each tissue type, the TCGA tissue IDs and the GEO Accessions for EPIC blood samples. For each tissue, we computed the variance distribution among the samples and selected two sets of CpGs: stable CpGs (**sCpGs**), showing variance values below the 1st percentile, and unstable CpGs (**uCpGs**), showing variance above the 99th percentile.

For the tumor samples, we collected CpGs from the dataset of informative differentially methylated CpGs (iDMCs) computed in[15]. Supplementary Table S3 (Supplementary File 1) reports the TCGA tissue IDs for each analyzed cancer type. This dataset was created starting from TCGA Illumina Infinium 450K assays. In particular, the authors selected CpGs whose methylation value was significantly different in the cancer tissue when compared to the normal counterpart or to a reference normal when the counterpart was not available. For each tissue the authors then selected the top 1000 differentially methylated CpGs having a variance value higher of a predefined threshold in cancer[16]. For each considered cancer type, we further divided these sets in hyper-methylated iDMC (hyper-iDMC) and hypo-methylated iDMC (hypo-iDMC) based on the observed trend of differential methylation[15].

Motif discovery.

Given a set of CpGs of interest (normal or iDMCs) and the whole set of CpGs probed in the Illumina Infinium 450K platform as a background set, we searched for enriched motifs by using the Discriminative Regular Expression Motif Elicitation (*DREME*) tool from the Multiple EM for Motif Elicitation (*MEME)* suite[17]. In particular, for each CpG set, we provided *DREME* with two fasta files containing the *hg19* 20 bp flanking sequence of the CpGs in the set of interest and of the CpGs in the background set, respectively.

From the *DREME* output, we selected only those motifs with an E-value, i.e. the enrichment Fisher p-value corrected for multiple test on the number of candidate motifs, lower than 0.05.

Where the information on differential methylation was available, we classified a motif as mostly hyper- (hypo-) methylated if more than 70% of the CpGs carrying the motif were hyper- (hypo-) methylated, in all other cases the motif methylation status was considered as undefined. Moreover, for each motif in each tissue, we performed a hypergeometric test to evaluate the enrichment of CpGs carrying the motif for hyper- or hypo-methylation with respect to the corresponding background set. We considered as significantly hyper- or hypo-methylated the motifs with hypergeometric p-value lower than 0.05.

Motifs genomic annotation.

For each discovered motif, we selected the set of CpGs containing the motif in their surrounding sequence. We then annotated each set with *hg19* genes information, using the *IlluminaHumanMethylation450kanno.ilmn12.hg19* R package. In particular, we counted the number of CpGs in the set related to gene regions (1stExon, 3′_UTR, 5′_UTR, Body, TSS1500, TSS200) and CGI regions (Island, N_Shelf, N_Shore, OpenSea, S_Shelf, S_Shore).

Motifs association with transcription factors.

In order to assess whether the discovered motifs encompass a Transcription Factor Binding Site (TFBS), we employed the *Tomtom* tool[18] from the *MEME* suite. For each motif, we used *Tomtom* to search against the JASPAR 2018 (*JASPAR2018_CORE_vertebrates_non_redundant)* sequence database provided by the *MEME* suite. We considered for subsequent analyses only Jaspar TFBS having at least 5 bp overlap with the motif sequence and an E-value lower than 10.

TF binding methylation affinity and expression.

For each motif we considered the set of TFs, derived from the *Tomtom* tool, showing binding affinity. We then considered only the transcription factors (TFs) annotated as MethylPlus or MethylMinus in[19]. For each TF, we retrieved the list of potential target genes using the TRRUST database[20]. For each motif/TF pair, we retrieved the expression profiles of the TF target genes in the associated tumor and normal tissues using the FirebrowseR package[16]. We then performed a test for differential expression using a Wilcoxon test and defined each gene upregulated or downregulated in cancer, if the difference in median expression values between tumor and control was positive or negative and the Wilcoxon p-value was lower than 0.05. After that, we annotated the motif/TF pairs as associated with up-regulation if more than 50% of differentially expressed target genes were up-regulated, down-regulation if more than 50% of differentially expressed target genes were down-regulated, mixed-regulated otherwise.

Results

Sequence context as a marker of methylation instability in normal tissues.

To investigate the relationship between DNA context and CpG methylation stability, we built a set of normal methylomes using DNA methylation data of normal tissues associated to 11 cancer types from TCGA and normal blood samples from the EPIC Italy cohort. Considering the variance distribution, we selected two sets of CpGs for each tissue: stable CpGs (**sCpGs**) and unstable CpGs (**uCpGs**).

For each sample, we considered in turn the set of selected stable and unstable CpGs and looked for recurrent DNA motifs (not necessarily including the CpG site) in the 20 bp flanking regions surrounding each CpGs and we found 108 motifs characterizing sCpGs and uCpGs in different tissues corresponding to overall 77 unique motifs (Supplementary Table S2). Out of these, 42 motifs were generated from uCpGs and 65 from sCpGs, with size ranging from 3 to 8 nucleotides (Supplementary Fig. S1).

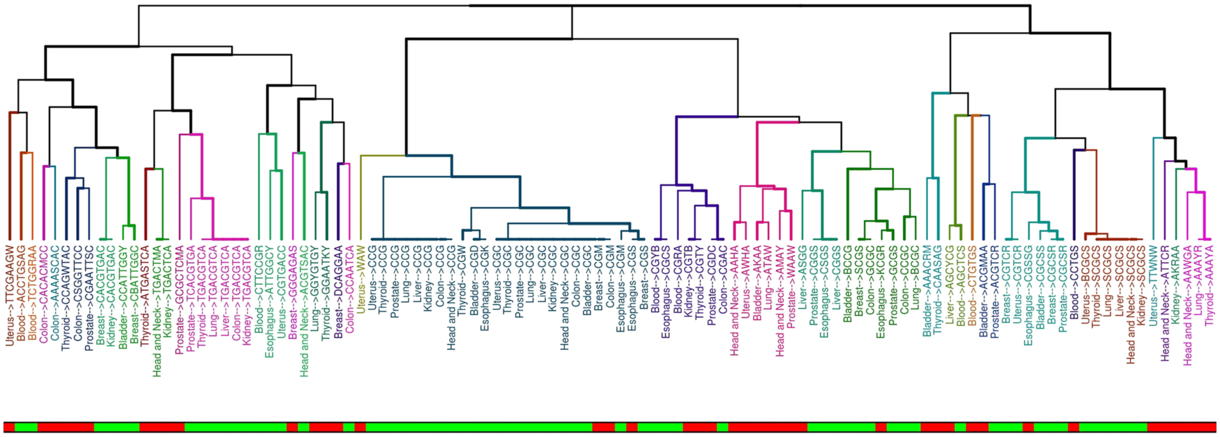

Figure 1. Clustering of motifs among the different normal tissues. For each motif the normal tissue and the sequence are reported. Labels' color represents groups of similar motifs based on clustering. The bottom bar indicates the stability of the obtained motifs. sCpGs are shown in green while uCpGs in red.

We then compared the discovered motifs among the different tissues by means of a distance function based on sequence similarity and grouped them using hierarchical clustering (Fig. 1).

Interestingly, we found distinct patterns of clustering among stable and unstable sequence motifs. By doing so, we were able to identify at least 9 groups of highly similar motifs that were recurrently found in more than 3 tissues.

Next, we characterized the relative positioning of motifs surrounding the stable and unstable CpGs by considering their localization with respect to genes (Fig. 2A) and to CpG dense regions (Fig. 2B). The sCpGs motifs are predominantly localized within the 5′ UTR and the 1st exon, while the uCpGs motifs are mainly localized within the gene body (Fig. 2A). Interestingly, although uCpGs motifs are primarily located within the openSea regions, they maintain a discrete distribution in other genomic locations (Fig. 2B). sCpGs motifs were mainly localized within the proximal gene regulatory regions as well as islands.

Sequence context as a marker of aberration in cancers.

We then considered 32 cancers from the TCGA (Supplementary Table S3) and used the dataset of informative differentially methylated CpGs (iDMCs) defined in[15]. We divided these sets, for each cancer type, in hyper-methylated (hyper-iDMC) and hypo-methylated iDMC (hypo-iDMC) based on their methylation trend in tumor with respect to normal. By following a similar approach as in the previous analysis we considered, for each cancer type, the entire set of iDMCs and looked for recurrent DNA motifs (not necessarily including the CpG sites) in the 20 bps flanking regions surrounding each iDMC and we discovered statistically overrepresented DNA motifs in almost all cancer types (Fig. 3, Supplementary Table S4). In particular, we found 120 motifs overall the considered tissues, corresponding to 114 unique motifs, with different sizes ranging from 3 to 8 nucleotides (Supplementary Fig. S2 and Table S4).

For each motif, we assigned a dominant methylation direction in cancer (Fig. 3, bottom bar). Considering the reported methylation status of the iDMCs carrying the motif, motifs were labeled as "hyper-methylated", "hypo-methylated" and "undefined" (see Methods section). This analysis led to 68 hypo-methylated, 36 hyper-methylated and 16 undefined motifs.

Moreover, we characterized the motifs to be significantly hyper- or hypo-methylated if the CpGs underlying the definition of the motif were significantly enriched (Hypergeometric test, $p < 0.05$), given the background composition of input iDMCs, for the particular aberration (Supplementary Table S2).

In order to ascertain how similar signatures were distributed among different cancer types and different aberration types (hypo- or hyper-methylation), we computed a distance function based on the sequence similarity among all pairs of derived motifs and performed a hierarchical clustering (Supplementary Fig. S3).

Moreover, we analyzed the localization of the motifs associated with aberrant methylation in cancer samples (Fig. 4). We observed a neat positioning pattern of hypermethylated motifs within CG dense regions (Fig. 4B). In fact, in BRCA, CESC, CHOL, COAD, ESCA, LGG, LUAD, PAAD, PRAD, STAD, UCEC and UCS, the hypermethylated motifs are mainly located within CG-rich regions as well as CpG islands, while a marked localization is visible within the shore regions in LGG and in UCEC. On the other hand, the hypomethylated motifs are located mostly in openSea regions. Concerning protein coding genes, we observed a predominance of hypomethylated regions falling within the gene bodies (Fig. 4A).

Binding site affinity in cancer motifs.

In order to investigate how the methylation status of enriched motifs can affect the transcriptional regulation of their potential target genes in cancer, we looked for transcription factors (TFs) whose binding domains have a potential affinity for the discovered motifs. Our analysis highlighted that most enriched transcription factors are both related to basal transcriptional regulation as well as to epigenetic mechanisms (Fig. 5). For instance, the Fos-Jun complex is the most enriched overall, followed by the Histone H4 nuclear factor P (HINFP). Similarly, we identified multiple motifs with potential binding affinity for the E2F transcription factors.

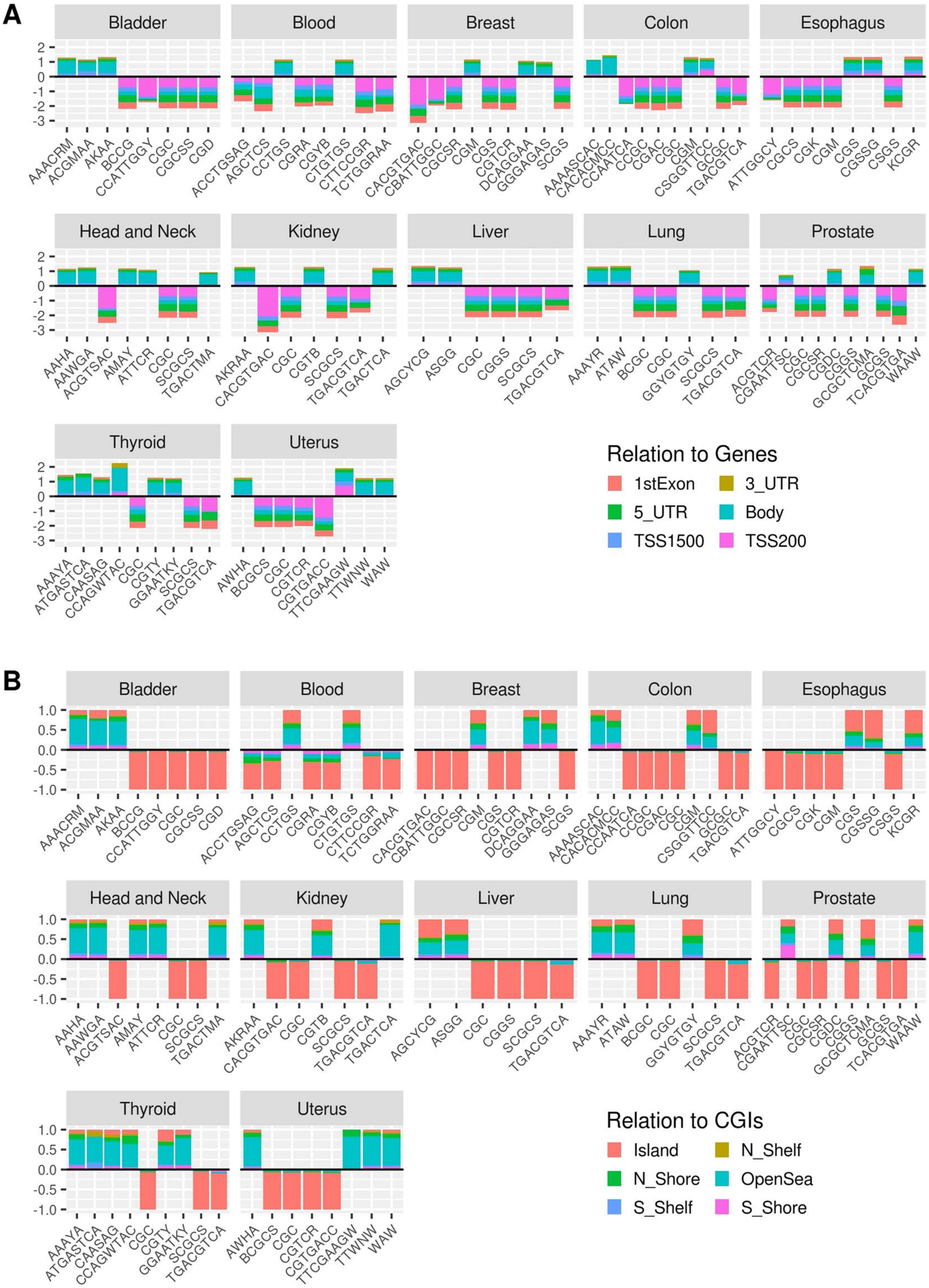

Figure 2. Genomic distribution of discovered motifs in normal tissues. Bar plots representing, for each analyzed tissue, the genomic distribution relative to gene regions (upper panel A) and CpG islands (lower panel B) respectively for motifs derived from uCpGs (positive values) and sCpGs (negative values).

To better characterize the regulatory role of the motifs in the malignant phenotype, we identified transcriptional regulators acting both in a specific cancer and across multiple types of cancers. As shown in Fig. 5, a plethora of motifs share affinity for common transcriptional regulators (even in different cancers), mainly involved in basal transcriptional regulation machinery, such as the above mentioned Fos-Jun heterodimer and their paralogue proteins, Bach and Maf.

Such regulators were enriched in several cancer types, as well as lung, bladder cancer and kidney renal clear cell carcinoma. On the other hand, a distinct regulatory signature sustained by the IRF family regulators shared by

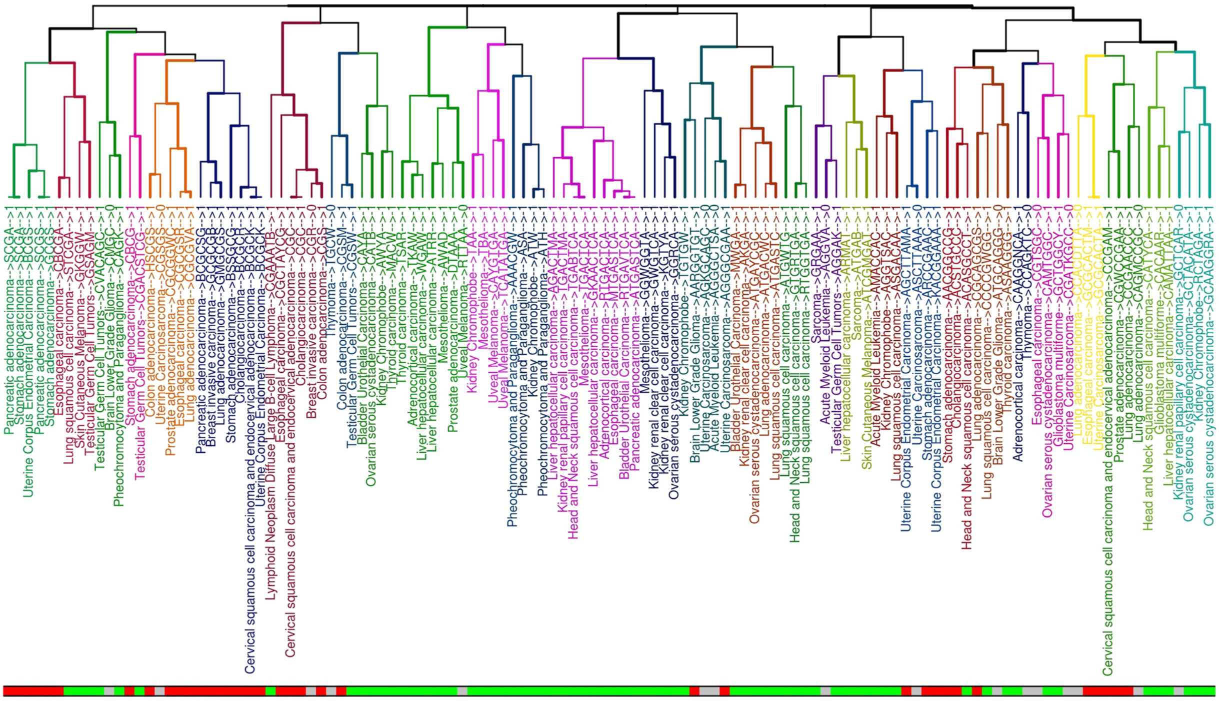

Figure 3. Clustering of motifs among different cancer tissues. For each motif the cancer tissue and the sequence are reported. The bottom bar indicates hypo-methylated motifs in green, hyper-methylated motifs in red, with undefined methylation status in grey. Labels' color represents groups of similar motifs based on clustering.

stomach, cervical squamous cell carcinoma and in lymphoid neoplasm diffuse large B-cell lymphoma emerged. Furthermore, our findings show a specific enrichment of the ETV family regulators in colon adenocarcinoma and uterine corpus endometrial carcinoma. While such patterns of regulation are peculiar for the mentioned cancers, we observed a remarkably complex regulation in lung and prostate adenocarcinoma as well as in esophageal cancer.

The methylation status of hyper-iDMC and hypo-iDMC affects gene expression through TFs binding. Recently, different affinity binding of TFs to their target sites based on their methylation status has been thoroughly characterized[19]. In order to strengthen our findings from a mechanistic point of view, we tested the hypothesis that there is a concordance between the methylation status of the considered motifs, the methylation-dependent binding affinity of the transcriptional regulators, and the expression trend of their target genes. To achieve this goal, we first classified each association TF/motif as concordant if the TF was reported as MethylPlus in[19] and associated to a hyper-methylated motif or if it was classified as MethylMinus and associated to a hypo-methylated motif in the corresponding cancer tissue, as discordant otherwise (Supplementary Table S5).

Summarizing over all cancer tissues, we found that when considering our motifs as potential binding sites, the concordance was confirmed only for MethylMinus TFs.

Then, we linked this information with differential expression of available cancer tissues and we found that a notable group of concordant TFs target genes show expression patterns in line with the methylation status of the motif and the TFs preference for methylated or unmethylated targets (Table 1).

Discussion

In this work, we explored the relationship between CpG methylation changes and the presence of recurring DNA sequences in the surrounding region of CpG sites. The analysis has been conducted on CpGs showing different methylation variation patterns in normal tissues and CpGs associated with aberrant methylation in cancer tissues.

In normal tissues, all of the motifs that were found enriched in more than one tissue, were predominantly associated to the same stability pattern. This suggests that their association with methylation stability is conserved across different tissues. Short motifs were mostly composed by a CpG dinucleotide followed by a further C or G, most probably indicating the origin of the motif in a CpG-rich region.

Interestingly, structures formed by more than 4 nucleotides, such as the *TGACGTCA* motif, were retrieved in their exact sequence in prostate, thyroid, lung, liver, colon and kidney. Notably, the CpG methylation status of such motif has been extensively studied. In fact, *TGACGTCA* has been identified as a pivotal factor in regulating cell type- and development stage-specific gene expression[21]. The *TGACGTCA* motif (namely the cAMP-responsive enhancer/promoter sequence) is specifically occurring upstream of genes whose expression is regulated by cyclic adenosine monophosphate molecules, and its aberrant methylation is related to the loss of binding properties by ATF family regulators and subsequent transcriptional inactivation.

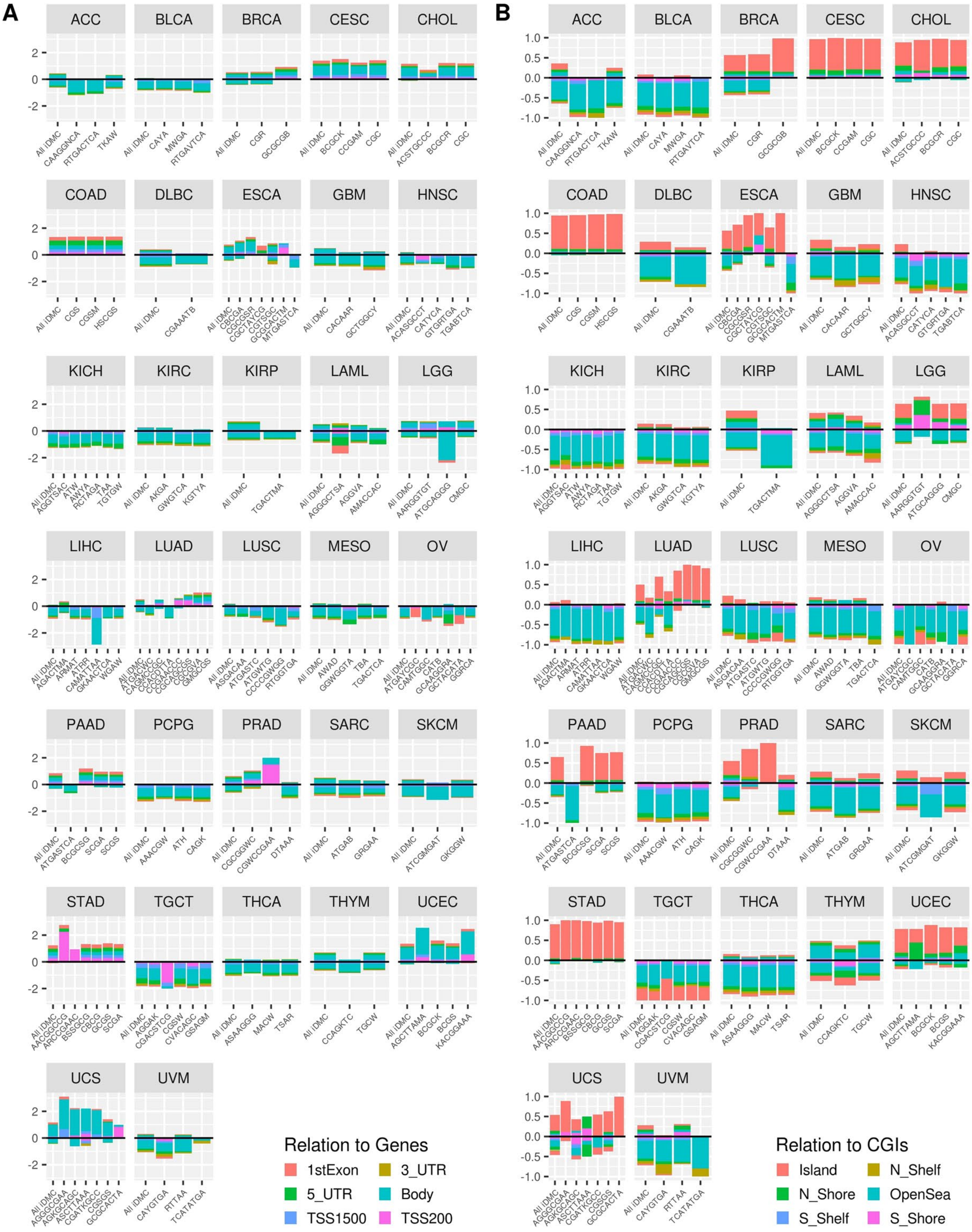

Figure 4. Genomic distribution of cancer methylation motifs. Bar plots representing, for each analyzed cancer tissue, the genomic distribution relative to gene regions (left panel A) and CpG islands (right panel B) respectively for motifs derived from hyper-iDMC (positive values) and hypo-iDMC (negative values).

Moreover, we found the *CACGTGAC* motif significantly overrepresented in breast and kidney. Anantharaman and colleagues[22] investigated the cis-regulatory role of this motif in the basal regulation of transcription highlighting the capability of such motif to bind HLH proteins in erythroid cell line differentiation. In this context, the *CACGTGAC* motif is known as E-box. Furthermore, early *in vitro* studies demonstrated that this motif constitutes the binding site for the transcription factors USF1 and USF2[23,24]. These transcriptional regulators can act both as homo or heterodimer recruiting the TFIID complex to the promoter. This motif has been already described as having a high affinity binding for multiple other transcription factors such as NF-E2, GATA-1, EKLF, and Tal1, *in vivo*[25].

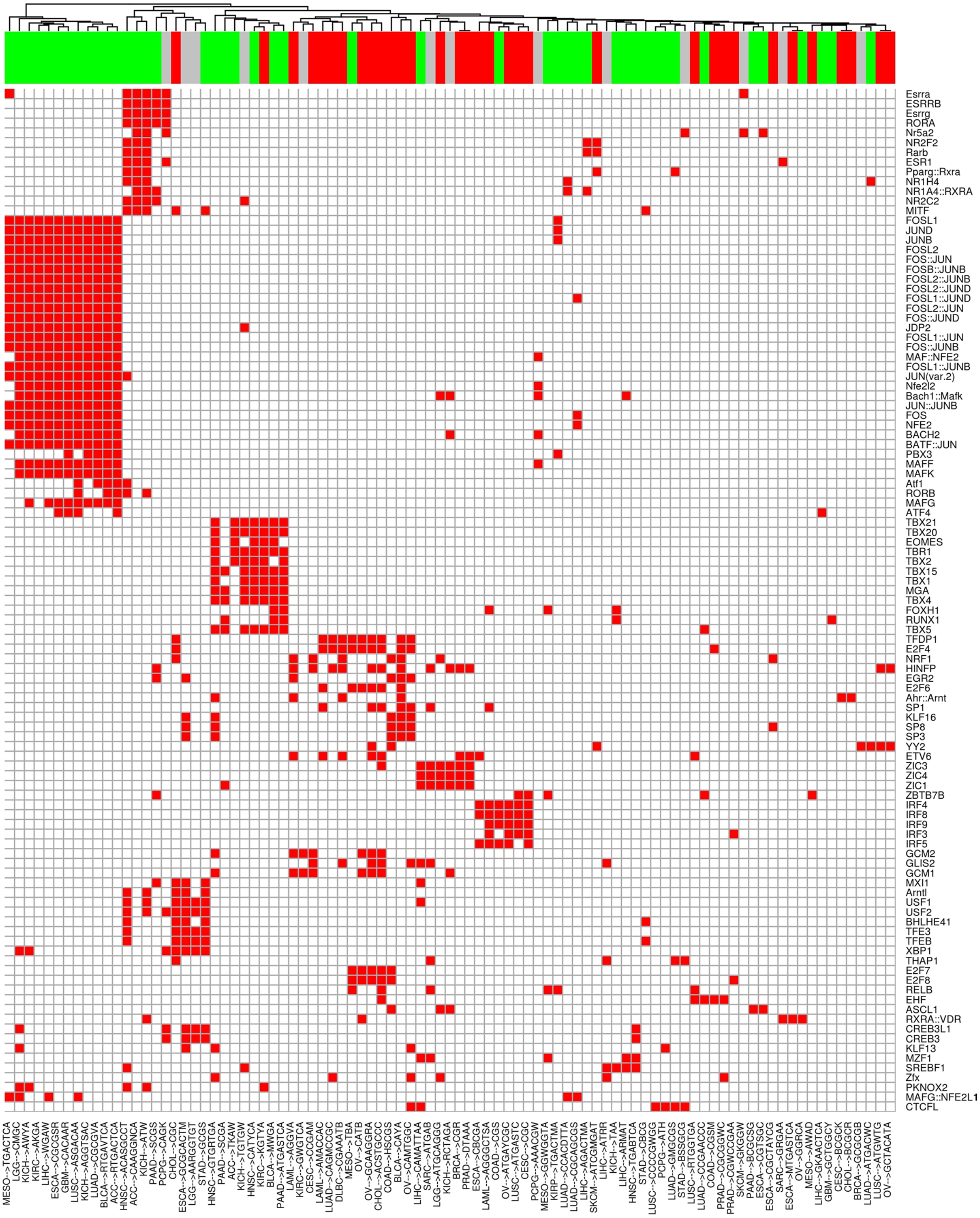

Figure 5. Transcription factors frequently associated with motifs. Heatmap reporting transcription factors whose binding sites are associated with motifs in at least 4 different tissues. Columns report each motif/cancer pair while rows report enriched TFs. Motifs are clustered based on the number of shared enriched TFs. The upper bar indicates hypo-methylated motifs in green, hyper-methylated motifs in red, with undefined methylation status in grey.

The *CCAATCA* motif, enriched in the colon tissue, is localized in the region surrounding one of the most diffused cis-regulators in the basal transcription machinery, the CCAAT box. Meergans and colleagues[26] clarified the role of this well conserved sequence in the expression regulation of genes coding for the histone proteins H1, the main factor guiding the nucleosome positioning. Interestingly, the so-called CCAAT box is highly conserved in all the promoters of the H1 loci and it is followed by a CA dinucleotide.

	Down-regulation	Mixed-regulation	Up-regulation
Concordant TFs	5	0	24
Discordant TFs	4	5	3

Table 1. Number of identified target genes showing only down-regulation, mixed regulation and up-regulation patterns among cancer tissues. Differentially expressed target genes among all cancer tissues are divided on the base of the concordance of their targeting TFs methylation affinity and their TFBS methylation status. We define as "Concordant TFs" the TFs classified as MethylPlus and associated to a hyper-methylated motif or classified as MethylMinus and associated to a hypo-methylated motif in the corresponding cancer tissue. On the contrary, we define as "Discordant TFs" the TFs classified as MethylMinus and associated to a hyper-methylated motif or classified as MethylPlus and associated to a hypo-methylated motif in the corresponding cancer tissue.

A correlation between the local methylation profiles and CpGs density is known[27–29]. CpGs proximal to a TSS show less variation (more stability) than distal CpGs[6]. In this context, our results support the hypothesis that stable CpG motifs are related to high CpG content and located in promoter regions, while unstable CpG motifs are related to low CpG content and located in gene body regions.

Similarly, in Acute Lymphoid Leukemia cells it has been shown that CpGs located outside the CGIs show a higher variation (instability) than those falling within the CGIs[30].

Methylation changes occur not only during physiological processes but also in several pathological conditions including cancer[25]. To date, the exact mechanism that modulates the sensitivity of CpGs to methylation is still unknown. A seminal study by Feltus *et al.* reported that aberrant methylation relies on local sequence composition[31].

Having observed the existence of motif structures associated with CpG methylation stability in normal tissues, we ascertained whether similar structures can be found when considering aberrant methylation associated with cancer tissues.

For instance, the *TGACTCA* motif, which was statistically overrepresented in mesothelioma and in its variant *TGACTMA* in kidney, has been already identified as the binding motif of the transcriptional complex AP1[32]. Seldeen *et al.*[32], inspected the effect of constitutive modifications of the *TGACTCA* motif, given the important role of the Fos-Jun transcription factors in the translation of extracellular signals in gene expression regulation (by growth factors and cytokines).

In order to disentangle the motif relatedness among different tissues from the similarity of the corresponding CpG input sets, we built a dendrogram based on the Jaccard distance among sets of iDMC for each tumor (Supplementary Fig. S3) and observed a different clustering of cancers, mainly driven by their tissue of origin. This suggests that tissue similarities found by considering motifs structure are not only dependent on the amount of shared iDMCs, but also by the structure composition of enriched motifs generated by different sets of CpG sites. When considering the methylation direction, we observed that the clustering of structurally similar motifs is consistent with the direction of methylation change. Notably, we observed that motifs which are prone to have a hypermethylated profile are GC-rich in respect to the other ones. This phenomenon has been previously described and reported in[33] as CpG island methylator phenotype (CIMP). Although CIMP has been reported in several cancers, including gastric[34–37], lung[38], liver[39], ovarian[40], endometrial[41], breast[42], gliomas[43] and leukemias[44], most of the literature relates its effects to colorectal cancers (CRC)[45]. Currently, the CIMP causative factors are still unknown[46]. However, the investigation of CIMP in different cancers led to the hypothesis of an epigenetic-driven onset of the malignant transformation, in turn leading to the inactivation of essential genes as well as tumor suppressors. The results obtained in the present work are also in line with this hypothesis. When looking at the genomic localization, we found that enriched motifs in STAD and PRAD are mainly localized within the 200 bp sequence surrounding the transcription start site, suggesting their possible direct involvement on the corresponding gene expression activity. Once obtained a list of motifs associated with each cancer type, we investigated their potential transcription factor binding affinity.

For instance, the motifs encompassing the HINFP binding sites, are among the most enriched overall the considered cancers. HINFP is a crucial regulator of the expression of genes encoding the H4 histones. Therefore, its function is essential in the S phase of the cell cycle, affecting the packaging of newly synthetized DNA into condensed chromatin[47]. This raises the perspective that aberrant methylation in motifs surrounding the HINFP binding sites could be causative of genomic instability and malignant transformation. On the other hand, our results suggest a possible involvement of the discovered motifs in the regulation of genes targeting the AP-1 heterodimer. This is also supported by the presence of motifs related to the NFE2 transcription factor, since this subunit is able to recognize the 5′-TGA(C/G)TCA-3′ sequence of the AP-1-like binding site[48]. This regulatory complex affects the expression of a wide range of genes. Dimerization of Fos-Jun proteins with members of the Maf and NF-E2 (CNC) families further expands the range of Fos-Jun targeting. In addition, different Jun and Fos family members can have opposite effects on AP-1-mediated transactivation[49–51]. Notably, by looking at the methylation status of the considered motifs, we noticed that this cluster of transcriptional regulators is typically related to hypomethylated motifs.

Similarly, markers of E2F-dependent transcription are mostly represented overall cancers types, as well as E2F4 and TFDP1, which regulate genes involved in key processes of malignant transformation, such as cell cycle regulation and DNA replication[52,53].

Deregulation of E2F-dependent G1/S transcription (i.e by certain oncogenes), leads the cells to enter the S-phase, triggering cell proliferation[54]. For this reason, we suppose that hypermethylation at E2F binding motifs, which we found overrepresented in colon and stomach adenocarcinoma as well as in sarcoma, could lead to an unmodulated E2F-dependent transcription of targets involved in the G1/S checkpoint and, thus, uncontrolled cell proliferation. Our results show that the TFs binding affinity for motifs marking aberrant DNA methylation regions involves both Fos-Jun regulators (as found in most of the cancers), and other specific factors, shared among few cancers, suggesting a possible crucial role of these motifs in the development of cancer phenotypes. On the opposite side, we identified motifs bound by a single transcription factor. In this case, the motif CGC, found overrepresented in cholangiocarcinoma, cervical squamous cell carcinoma and endocervical adenocarcinoma, is predicted to be bound specifically by Ahr::Arnt. Similarly, the motifs CGR (breast invasive carcinoma), CACAAR (glioblastoma multiforme), ATGWTG (lung squamous cell carcinoma), CGWCCGAA (prostate adenocarcinoma) show exclusive affinity for YY2, RUNX1, ATF4 and ZBTB7B, respectively.

We finally checked the concordance among the methylation status of the observed motifs, the TFs binding affinity for methylated or unmethylated sites and TFs target genes expression deregulation. The observed results suggest that the identified DNA methylation motifs not only represent potential biomarkers, but they contribute to disentangle the complex regulatory circuits in cancer genomes.

Conclusions

The relationship between sequence context and DNA methylation is an important feature to gain further insights on the dynamics of methylation establishment, maintenance and alteration in pathological conditions. Here, we provided the first comprehensive catalogue of short DNA sequences that can be associated to methylation stability between individuals in normal tissues and to aberrant methylation conditions in different cancer types. We showed how different DNA contexts could predispose a CpG to be more or less prone in gaining or losing methyl groups. We characterized the genomic localization of these sequences and their relationship with transcriptional regulators in the genome. When considering motifs associated to aberrant CpGs characterizing cancer tissues, we showed that some motifs are shared by different cancer types while others are cancer specific. This latter aspect could be exploited to define more precise and effective molecular targets in gene-based therapies such as those employing epigenetic CRISPR-Cas9 strategies.

Received: 3 May 2019; Accepted: 13 January 2020;
Published online: 03 February 2020

References

1. Szyf, M. Nongenetic inheritance and transgenerational epigenetics. *Trends. Mol. Med*. Elsevier Ltd. **21**, 134–44 (2015).
2. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: From the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).
3. Boyes, J. & Bird, A. Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.* **11**, 327–333 (1992).
4. Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Curr. Opin. Genet. Dev.* **25**, 85–92 (2014).
5. Weber, M. & Schübeler, D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell. Biol.* **19**, 273–280 (2007).
6. Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* **15**, R37 (2014).
7. Haerter, J. O., Lövkvist, C., Dodd, I. B. & Sneppen, K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. *Nucleic Acids Res.* **42**, 2235–2244 (2014).
8. McCabe, M. T., Brandes, J. C. & Vertino, P. M. Cancer DNA Methylation: Molecular Mechanisms and Clinical Implications. *Clin. Cancer Res.* **15**, 3927–3937 (2009).
9. Bartlett, T. E.*et al.* Corruption of the Intra-Gene DNA Methylation Architecture Is a Hallmark of Cancer. *PLoS One* **8** (2013).
10. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
11. Ghorbani, M., Themis, M. & Payne, A. Genome wide classification and characterisation of CpG sites in cancer and normal cells. *Comput. Biol. Med.* **68**, 57–66 (2016).
12. Vaiopoulos, A. G., Athanasoula, K. C. & Papavassiliou, A. G. Epigenetic modifications in colorectal cancer: Molecular insights and therapeutic challenges. *Biochim. Biophys. Acta* **1842**, 971–980 (2014).
13. Cancer Genome Atlas Research Network *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*. NIH Public Access; **45**, 1113–1120, http://www.ncbi.nlm.nih.gov/pubmed/24071849 (2013).
14. Calza, S. *et al.* EPIC-Italy cohorts and multipurpose national surveys. A comparison of some socio-demographic and life-style characteristics. **89**, 615–623, http://www.ncbi.nlm.nih.gov/pubmed/14870826.
15. Zheng, X., Zhang, N., Wu, H. J. & Wu, H. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* **18**, 17, https://doi.org/10.1186/s13059-016-1143-5 (2017).
16. Deng, M., Brägelmann, J., Kryukov, I., Saraiva-Agostinho, N. & Perner, S. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database*, http://www.ncbi.nlm.nih.gov/pubmed/28062517 (2017).
17. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659, http://www.ncbi.nlm.nih.gov/pubmed/21543442 (2011).
18. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. Quantifying similarity between motifs. *Genome Biol.* **8**, R24, https://doi.org/10.1186/gb-2007-8-2-r24 (2007).
19. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239, http://www.ncbi.nlm.nih.gov/pubmed/28473536 (2017).
20. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–386, http://www.ncbi.nlm.nih.gov/pubmed/29087512 (2018).
21. Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev*. **3**, 612–619, http://www.ncbi.nlm.nih.gov/pubmed/2545524 (1989).
22. Anantharaman, A. *et al.* Role of Helix-Loop-Helix Proteins during Differentiation of Erythroid Cells. *Mol. Cell Biol.* **31**, 1332–1343, http://www.ncbi.nlm.nih.gov/pubmed/21282467 (2011).

23. Sawadogo, M., Van Dyke, M. W., Gregor, P. D. & Roeder, R. G. Multiple forms of the human gene-specific transcription factor USF. I. Complete purification and identification of USF from HeLa cell nuclei. *J. Biol. Chem.* **263**, 11985–11993, http://www.ncbi.nlm.nih.gov/pubmed/3403558 (1988).
24. Sawadogo, M. Multiple forms of the human gene-specific transcription factor USF. II. DNA binding properties and transcriptional activity of the purified HeLa USF. *J. Biol. Chem.* **263**, 11994–12001, http://www.ncbi.nlm.nih.gov/pubmed/3403559 (1988).
25. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2**, S4–11 (2005).
26. Meergans, T., Albig, W. & Doenecke, D. Conserved sequence elements in human main type-H1 histone gene promoters: their role in H1 gene expression. *Eur. J. Biochem.* **256**:436–46, http://www.ncbi.nlm.nih.gov/pubmed/9760185 (1998).
27. Lam, L. L. *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proc. Natl. Acad. Sci.* **109**, 17253–17260 (2012).
28. Jiang, R. *et al.* Discordance of DNA Methylation Variance Between two Accessible Human Tissues. *Sci. Rep.* **5**, 8257 (2015).
29. Palumbo, D., Affinito, O., Monticelli, A. & Cocozza, S. DNA Methylation variability among individuals is related to CpGs cluster density and evolutionary signatures. *BMC Genomics.* **19**, 229 (2018).
30. Milani, L. *et al.* DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Child A. Glob. J. Child Res.* **115**, 1214–1225 (2010).
31. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. DNA motifs associated with aberrant CpG island methylation. *Genomics.* **87**, 572–579, http://www.ncbi.nlm.nih.gov/pubmed/16487676 (2006).
32. Seldeen, K. L., McDonald, C. B., Deegan, B. J. & Farooq, A. Single nucleotide variants of the TGACTCA motif modulate energetics and orientation of binding of the Jun-Fos heterodimeric transcription factor. *Biochemistry.* **48**, 1975–1983, http://www.ncbi.nlm.nih.gov/pubmed/19215067 (2009).
33. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proc. Natl Acad. Sci. USA* **96**, 8681–8686 (1999).
34. An, C. *et al.* Prognostic significance of CpG island methylator phenotype and microsatellite instability in gastric carcinoma. *Clin. Cancer Res.* **11**, 656–663 (2005).
35. Kusano, M. *et al.* Genetic, epigenetic, and clinicopathologic features of gastric carcinomas with the CpG island methylator phenotype and an association with Epstein–Barr virus. *Cancer.* **106**, 1467–1479 (2006).
36. Oue, N. *et al.* DNA methylation of multiple genes in gastric carcinoma: association with histological type and CpG island methylator phenotype. *Cancer Sci.* **94**, 901–905 (2003).
37. Toyota, M. *et al.* Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. *Cancer Res.* **59**, 5438–5442 (1999).
38. Marsit, C. J. *et al.* Examination of a CpG island methylator phenotype and implications of methylation profiles in solid tumors. *Cancer Res.* **66**, 10621–10629 (2006).
39. Shen, L. *et al.* DNA methylation and environmental exposures in human hepatocellular carcinoma. *J. Natl. Cancer Inst.* **94**, 755–761 (2002).
40. Strathdee, G. *et al.* Primary ovarian carcinomas display multiple methylator phenotypes involving known tumor suppressor genes. *Am J. Pathol.* **158**, 1121–1127 (2001).
41. Sasaki, M. *et al.* Multiple promoters of catechol-O-methyltransferase gene are selectively inactivated by CpG hypermethylation in endometrial cancer. *Cancer Res.* **63**, 3101–3106 (2003).
42. Fang, F. *et al.* Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.* **3**, 75ra25 (2011).
43. Li, Q. *et al.* Concordant methylation of the ER and N33 genes in glioblastoma multiforme. *Oncogene* **16**, 3197–3202 (1998).
44. Garcia-Manero, G. *et al.* DNA methylation of multiple promoter-associated CpG islands in adult acute lymphocytic leukemia. *Clin. Cancer Res.* **8**, 2217–2224 (2002).
45. Hughes, L. A. E. *et al.* The CpG island methylator phenotype in colorectal cancer: Progress and problems. *Biochimica et Biophysica Acta* **1825**, 77–85 (2012).
46. Nazemalhosseini, M. E., Kuppen, P. J., Aghdaei, H. A. & Zali, M. R. The CpG island methylator phenotype (CIMP) in colorectal cancer. *Gastroenterol Hepatol Bed Bench.* **6**, 120–128 (2013).
47. Ghule, P. N. *et al.* p53 checkpoint ablation exacerbates the phenotype of Hinfp dependent histone H4 deficiency. *Cell Cycle.* **14**, 2501–2508, http://www.ncbi.nlm.nih.gov/pubmed/26030398 (2015).
48. Daftari, P., Gavva, N. R. & Shen, C. K. J. Distinction between AP1 and NF-E2 factor-binding at specific chromatin regions in mammalian cells. *Oncogene.* **18**, 5482–5486, http://www.ncbi.nlm.nih.gov/pubmed/10498903 (1999).
49. Chen, T. K., Smith, L. M., Gebhardt, D. K., Birrer, M. J. & Brown, P. H. Activation and inhibition of the AP-1 complex in human breast cancer cells. *Mol. Carcinog.* **15**, 215–226, http://www.ncbi.nlm.nih.gov/pubmed/8597534 (1996).
50. Kharman-Biz, A. *et al.* Expression of activator protein-1 (AP-1) family members in breast cancer. *BMC Cancer* **13**, 441, http://www.ncbi.nlm.nih.gov/pubmed/24073962 (2013).
51. Ding, X. *et al.* Epigenetic Activation of AP1 Promotes Squamous Cell Carcinoma Metastasis. *Sci. Signal.* **6**, ra28–ra28, http://www.ncbi.nlm.nih.gov/pubmed/23633675 (2013).
52. Bertoli, C., Herlihy, A. E., Pennycook, B. R., Kriston-Vizi, J. & de Bruin, R. A. M. Sustained E2F-Dependent Transcription Is a Key Mechanism to Prevent Replication-Stress-Induced DNA Damage. *Cell. Rep.* **15**, 1412–1422, http://www.ncbi.nlm.nih.gov/pubmed/27160911 (2016).
53. Hsu, J. & Sage, J. Novel functions for the transcription factor E2F4 in development and disease. *Cell Cycle.* **15**, 3183–3190, http://www.ncbi.nlm.nih.gov/pubmed/27753528 (2016).
54. Bertoli, C., Skotheim, J. M. & de Bruin, R. A. M. Control of cell cycle transcription during G1 and S phases. *Nat. Rev. Mol. Cell. Biol.* **14**, 518–528, http://www.ncbi.nlm.nih.gov/pubmed/23877564 (2013).

Acknowledgements

This study was supported by the Academy of Finland (grant agreements 275151 and 292307).

Author contributions

G.S. and D.G. conceived the concept and supervised the study; G.S. and D.P. carried out the bioinformatic analyses; G.S., D.G., S.C., D.P. and A.F. interpreted and commented the results; G.S., D.G., S.C., D.P. and A.F. drafted the manuscript. All the authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58331-w.

Correspondence and requests for materials should be addressed to D.G.

Reprints and permissions information is available at www.nature.com/reprints.