Nucleic Acids Research, 2025, **53**, gkaf949
https://doi.org/10.1093/nar/gkaf949
Computational Biology

Systematic identification of regions where DNA methylation is correlated with transcription refines regulatory logic in normal and tumour tissues

Richard Heery* **and Martin H. Schaefer** ***

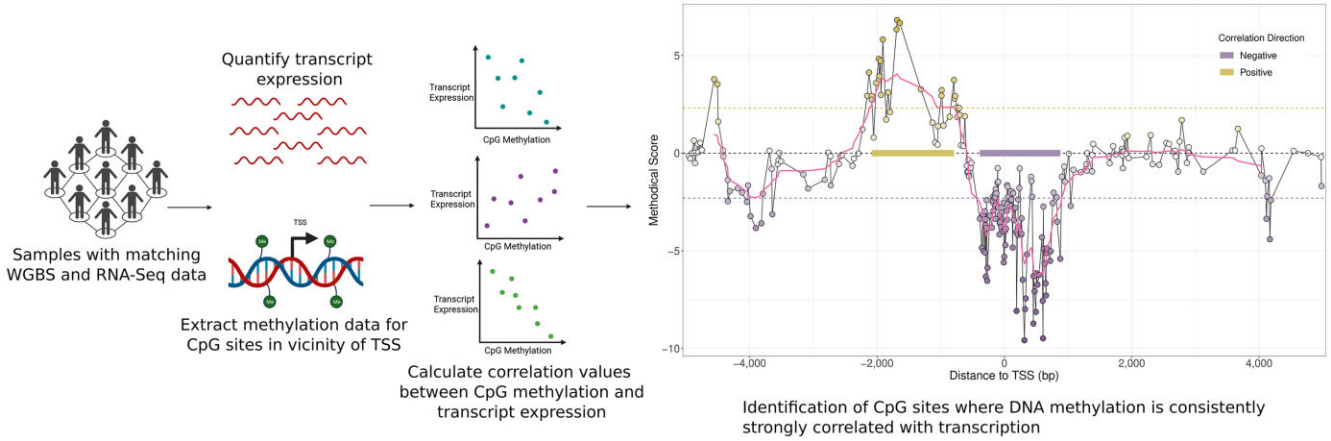
Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Via Adamello 16, 20139 Milan, Italy

*To whom correspondence should be addressed. Email: martin.schaefer@ieo.it
Correspondence may also be addressed to Richard Heery. Email: richardheery@gmail.com

Abstract

DNA methylation at gene promoters is generally considered to be associated with transcriptional repression in vertebrates. However, lack of a clear picture of where promoter methylation is most important for transcriptional regulation has hindered our understanding of this relationship and resulted in the use of a wide variety of arbitrary promoter definitions. We demonstrate here that the use of different promoter definitions can lead to contradictory results between studies of promoter methylation. In response, we have developed Methodical, a computational method that combines RNA-seq and whole genome bisulfite sequencing (WGBS) data to identify genomic regions where DNA methylation is highly correlated with transcriptional activity. We refer to these regions as transcript-proximal methylation-associated regulatory sites (TMRs). We applied Methodical to one normal prostate tissue data set, one prostate tumour dataset, and one prostate metastasis dataset and characterized the identified TMRs. We show that the region just downstream of the TSS is the most common location for TMRs and that TMRs are enriched for particular genomic features, chromatin states, and transcription factor binding sites. Finally, we demonstrate that the methylation of TMRs is generally strongly correlated with transcription in diverse cancer types and that TMRs are highly subject to altered DNA methylation in cancer.

Graphical abstract



Introduction

Gene promoters are broadly defined as the region in the vicinity of transcription start sites (TSS) where RNA polymerase is recruited by transcription factors to initiate transcription. DNA methylation involves the addition of methyl groups to DNA and in mammals generally occurs at the 5′ position of cytosine bases that are followed by a guanine base (CpG sites) [1–3]. It is generally accepted that DNA methylation at gene promoters has a repressive effect on transcription [3–5]. Two main models have been put forward to describe how DNA methylation could mediate this repression. The first model proposes that transcription factor binding is blocked indirectly via recruitment of methylated DNA-specific proteins, which, in turn, recruit co-repressors that silence transcription, while the second proposes that DNA methylation directly blocks transcription factor binding [4].

However, genome-wide studies have reported fairly weak correlations in general between promoter DNA methylation and gene expression [6–8]. Additionally, exceptions to the general paradigm of gene silencing by DNA methylation have been reported where DNA methylation has been found to be positively associated with the expression of certain genes [9, 10]. Thus, the relationship between DNA methylation and transcriptional activity seems to be more nuanced and

Received: February 1, 2024. **Revised:** July 12, 2025. **Accepted:** August 23, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

context-dependent than suggested by the conventional model. Moreover, DNA methylation at regions downstream of the TSS, particularly the first intron and exon, has been recognized to be associated with transcriptional silencing [11, 12]. Additionally, variably methylated regions outside promoter regions were found to often display stronger correlations with gene expression than promoter regions when studying single cells [13].

When studying promoter DNA methylation, typically CpG methylation values have been averaged over the extent of the designated promoter region and linear models then fit to gene expression values. Several recent studies have shown that this simple approach often fails to accurately capture the relationship between DNA methylation and transcription, with more sophisticated models incorporating distal regions and considering methylation at a higher resolution providing far more accurate predictions of transcriptional activity. In particular, a probabilistic machine learning approach that extracted higher order methylation features produced a very powerful predictor of gene expression [14]. Alternative machine learning approaches have also provided further insight into the association between DNA methylation and transcription, highlighting the importance of the region just downstream of the TSS as well as distal regions [15–17]. Instead of directly predicting gene expression, another approach sought to predict promoter activity indirectly by using the enrichment of H3K4me3 and H3K27ac around the TSS as a proxy and predicting their levels using DNA methylation. The predicted levels of these marks in turn were found to be relatively well correlated with transcriptional activity as measured by RNA-seq [18].

Another major limitation of previous studies of promoter DNA methylation has been the lack of agreement on the extent of promoters around TSS, resulting in the use of a wide variety of completely arbitrary promoter definitions. These different definitions have varied from hundreds to thousands of base pairs in length and also differed in the proportion of the promoter sequence located downstream of the TSS relative to upstream [19–23]. Indeed, it is difficult to find two studies using precisely the same promoter definition.

Variation in choice of promoter definition could obviously lead to inconsistent results between studies of promoter DNA methylation, including those aiming to identify promoters affected by DNA methylation change in cancer. Additionally, alternative promoter definitions could lead to the calculation of different correlation values between promoter methylation and transcript expression, obscuring our understanding of this relationship and also the recognition of which methylation changes in cancer are associated with corresponding transcriptional changes.

Most large-scale studies of DNA methylation over the last decade and a half have utilized the Illumina methylation microarrays due to their relative cost-effectiveness. This includes projects studying molecular alterations in diverse cancer types such as TCGA [24] and TARGET [25]. However, these microarrays measure DNA methylation only at a small percentage of the over 29.4 million CpG sites present in the human reference genome. The Infinium HumanMethylation450 array measures methylation at about 450 000 CpG sites, 1.5% of the total number present in the genome, and mostly targets CpGs located in CpG islands and promoters. The most recent generation of Illumina methylation microarray, the Infinium MethylationEPIC v2.0 array, targets over 935 000 CpG sites and has expanded the coverage of other regions compared to the HumanMethylation450 array, particularly of enhancers. It still, however, only covers about 3% of human CpG sites, and the targeted CpG sites are still highly enriched for certain genomic regions, such as CpG islands [26]. Thus, most attempts to study the relationship between DNA methylation and transcription across the genome have been limited to a small proportion of CpG sites and biased towards certain genomic contexts [27, 28].

The publication over the last few years of datasets with whole genome bisulfite sequencing (WGBS) and RNA-seq data for large numbers of human samples provide the opportunity to search for regions where DNA methylation is highly correlated with transcription in an unbiased manner. Thus, we developed Methodical, an algorithm that systematically identifies regions displaying such correlations. Two of the largest datasets to date have been for prostate cancer: one for prostate tumours and matching normal prostate samples [21] and another for prostate metastases [29]. We divided the prostate tumour and matching normal prostate dataset into two separate datasets for prostate tumours and normal prostate samples in order to study the relationship between DNA methylation and transcription in cancer and normal tissues separately. We applied Methodical to the three different datasets and subsequently characterized the identified regions, revealing novel insights into the relationship between DNA methylation and transcription in both normal prostate tissue and prostate cancer.

Materials and methods

Datasets

We used three different datasets to identify TMRs: one for normal prostate with 126 samples, one for prostate tumours with 126 samples, and one for prostate metastases with 99 samples. The normal prostate and prostate tumour samples come from the CPGEA project [21], and the prostate metastasis samples come from the MCRPC project [29].

Location of TSS and transcript-encoding regions

The location of TSS and the associated transcript-encoding regions for all protein-coding transcripts were obtained from Gencode (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gtf.gz). The genomic location of the first base of each transcript was designated as a TSS. To identify a subset of high-confidence TSS from among these, we used data produced by cap analysis of gene expression (CAGE), a technique to profile the 5′ ends of mRNA molecules [30], from the FANTOM5 project [31, 32].

We downloaded CAGE data for human prostate (http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.tissue.hCAGE/prostate%252c%2520adult%252c%2520pool1.CNhs10628.10022-101D4.hg19.ctss.bed.gz) and filtered for TSS supported by at least 10 CAGE tags. This gave us a set of 17 071 high-confidence TSS for 10 027 different genes. We used this set of TSS for all analyses unless stated otherwise.

Transcript expression

Transcript sequences were downloaded from Gencode (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.transcripts.fa.gz). The expression of

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

transcripts was quantified from paired end RNA-Seq FASTQ files using Kallisto (version 0.46.1) [33] with 100 bootstrap samples. Transcript counts were then normalized using the median of ratios approach from DESeq2 (version 1.46.0) [34].

DNA methylation–transcription correlations and methylation of genomic regions

All correlation values mentioned are Spearman correlation values. Statistical significance of correlation values was calculated by using the pt() function in R with the t-statistics derived from the correlation values. When calculating the methylation values of genomic regions, the mean of the methylation values of all CpG sites overlapping them was used. Differential methylation for genomic regions was tested using the diff_methylsig() function from the methylSig R package [35] (version 1.18.0) using the number of methylated reads and total number of WGBS reads overlapping each region.

Gene set overrepresentation analysis

MSigDB Hallmark pathways and KEGG pathways were obtained from MSigDB [36, 37] version 7.5 using the msigdb R package (version 1.6.0). Pathway overrepresentation among TMR-associated genes was performed by taking all genes associated with a given group of TMRs and testing the significance of their overlap with gene sets using the one-sided version of Fisher's exact test using all protein-coding genes as the background. *P*-values were adjusted using the Benjamini–Hochberg procedure.

Genomic regulatory features and chromatin states

The locations of CpG islands, CpG shores, CpG shelves, and open sea regions were obtained using the annotatr R package [38] (version 1.32.0). Predicted promoter regions, predicted enhancer regions, open chromatin regions and CCCTC-binding factor (CTCF) binding sites were downloaded from Ensembl version 114 using the biomaRt R package [39] (version 2.62.1). Chromatin states for prostate tissue were downloaded from https://download.cncb.ac.cn/OMIX/OMIX237/OMIX237-64-02.zip and lifted over to hg38 using the rtracklayer R package [40] (version 1.66.0).

To test if these regions were enriched among TMRs, we randomly shuffled the locations of TMRs within the search region (transcript-encoding regions with 5 kb added upstream and downstream) 1000 times using the regioneR package [41] (version 1.38.0). We then compared the observed overlap between the TMRs and the genomic regions of interest (measured in base pairs) to the distribution of overlap sizes for the random permutations to calculate relative enrichment and associated *P*-values. *P*-values were adjusted using the Benjamini–Hochberg procedure.

Transcriptional regulator binding site enrichment

The locations of binding sites for transcriptional regulators were obtained from ReMap2022 [42]. The BED file for non-redundant peaks in human for hg38 was downloaded from https://remap.univ-amu.fr/storage/remap2022/hg38/MACS2/remap2022_nr_macs2_hg38_v1_0.bed.gz and filtered for any binding sites identified in primary prostate tissue or prostate tumours or else cell lines derived from prostate tissue or prostate cancer (22Rv1, DU145, DuCAP, LAPC-4, LNCaP, LNCaP-95, LNCaP-abl, LNCaP-C4-2, LNCaP-C4-2B, LNCaP-FGC, MDA-PCa-2b, PC-3, RWPE-1, RWPE-2, VCaP, VCaP-LTAD, and WPMY-1). There were 77 different transcriptional regulators that had binding sites in prostate tissue, tumours, or cell lines.

The enrichment of binding sites among TMR groups was tested by comparing the proportion of CpG sites within TMRs which overlap binding sites with that of all CpGs within the search region for TMRs (transcript-encoding regions with 5 kb added upstream and downstream) using a two-sided Chi-squared test. *P*-values were adjusted using the Benjamini–Hochberg procedure.

Methodical algorithm

Spearman correlation values are calculated between the expression of a given transcript and the methylation values of all CpG sites within a specified region surrounding the TSS. We masked all CpG sites with less than 10 WGBS reads covering them and required at least 30 samples with non-missing methylation data at a given CpG site in order to calculate a correlation value and test its significance. We used the entire transcript-encoding region plus 5 kb upstream of the TSS and 5 kb downstream of the transcription end site (TES) of each transcript. The significance of these correlations is inferred and the resulting *P*-values are corrected for multiple testing using the Benjamini–Hochberg procedure. These corrected *P*-values are then transformed by taking their logarithm to the base 10 and multiplying by −1 if the correlations are positive, giving what we term the Methodical scores for the correlations. Thus, positive correlations have positive Methodical scores and negative correlations have negative Methodical scores, with the magnitude of the scores determined by the statistical significance of the correlations.

Smoothing of data across CpG sites has previously been employed to reduce noise in WGBS data [43], and Methodical adapts that approach by smoothing scores using an exponential moving average. The moving average employs a window that is centred on a given CpG site and includes an equal number of flanking CpGs upstream and downstream of this central CpG. For example, with five flanking CpGs, there would be a window size of 11 consisting of the central CpG, five CpGs upstream and five CpGs downstream. Weights decay geometrically and symmetrically moving away from the central CpG across the other CpGs in the window using a specified smoothing factor.

Two symmetrical significance thresholds are used to identify negative and positive TMRs: A Methodical score of $\log_{10}(0.05)$ is used to identify negative TMRs and a Methodical score of $-\log_{10}(0.05)$ to identify positive TMRs. Wherever the smoothed Methodical scores exceeds one of these thresholds, we group all CpGs within that region as a TMR of the corresponding direction.

We assessed different combinations of the smoothing factor for the exponential moving average and the number of flanking CpGs used to construct windows. We then evaluated the TMRs identified in the metastasis samples by calculating the correlations between methylation of the identified TMRs and expression of their associated transcript in prostate tumour samples. This approach allowed us to gauge the reproducibility of TMRs across different datasets. We found that a smoothing factor of 0.75 and windows constructed using 10 flanking CpGs resulted in a good trade-

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

off between the number of TMRs identified and their reproducibility (Supplementary Fig. S1A and B).

Refinement of Methodical

We investigated if the number of CpGs contained by TMRs was associated with the correlation between their methylation level and transcription of their associated transcript, which we will henceforth refer to just as the TMR correlation values. We discovered that there was a significant correlation between the number of CpGs TMRs contained and the strength of their correlation values (Supplementary Fig. S2). We thus decided to filter TMRs for those containing at least 5 CpG sites.

We discovered a huge number of positive TMRs in the prostate tumour and metastasis samples, particularly beyond a few kb from the TSS where they became far more common than negative TMRs (Supplementary Fig. S3). Attempts to correct for tumour purity did not change this. We suspected instead that this may be related to mappability, the varying ability to uniquely map reads across the genome. The conversion of unmethylated cytosines to thymines during WGBS reduces the overall mappability of the genome [44]. We thus hypothesized that the huge number of positive TMRs in cancer samples were possibly related to difficulty mapping WGBS reads to regions with low mappability, especially repetitive regions, leading to alignment ambiguity and inaccurate methylation calls in these regions. Using the WGBS mappability track for the Bismap project [44] (https://bismap.hoffmanlab.org/raw/hg38/k100.bismap.bedgraph.gz), we took a conservative approach of defining regions with a mappability score less than 1 as poorly mappable. We subsequently examined the distribution of TMRs overlapping regions with poor mappability (Supplementary Fig. S4A), revealing that a large proportion of the positive TMRs overlapped these regions.

We decided to keep only TMRs, which overlapped regions with the highest mappability as we felt we could not be confident in the validity of the rest. Afterwards, we found that the vast majority of positive TMRs located at a large distance from the TSS were removed, leaving a clear large peak of negative TMRs close to the TSS (Supplementary Fig. S4B).

We finally evaluated how the number of TMRs identified varies with the number of samples by counting the number of TMRs identified when running Methodical with differently sized subsets of normal prostate samples. We found that with less than 60 samples, few TMRs are identified (Supplementary Fig. S5), hinting that a minimum cohort size is required for a useful application of Methodical.

Software environment

All analyses were carried out using R version 4.4.3 on an Ubuntu 24.04 machine.

Results

Variable promoter definitions lead to inconsistent differential methylation results

Studies of promoter DNA methylation have employed a variety of different promoter definitions. Fig. 1A shows five different promoter definitions used in five different published studies of DNA methylation in cancer (A(19), B(20), C(21), D(22), and E(23)). Use of such different promoter definitions could obviously result in different promoter methylation levels being calculated for the same gene or transcript, potentially having a drastic impact on differential methylation results. To investigate this potential problem thoroughly, we decided to compare the differential promoter methylation results obtained using different promoter definitions with the same data set. Thus, we evaluated differential promoter methylation testing in prostate tumour samples compared to matching normal prostate samples using each of the five promoter definitions from Fig. 1A with a prostate cancer WGBS dataset [21].

Strikingly, we observed that different promoter definitions can lead to completely opposing results for the same TSS. Fig. 1B and C show the effect of the choice of promoter definition on differential promoter methylation for the genes *FLT1*, which encodes the vascular endothelial growth factor receptor 1 and which has been reported as hypermethylated in prostate cancer [45], and *POLR2H*, encoding one of the RNA polymerase subunits. We found that their promoters were identified as hypermethylated when using one set of promoter definitions but hypomethylated when using another set. We noted that there can be over a thousand TSS where a given pair of different promoter definitions lead to such contradictory results (Supplementary Fig. S6).

We also observed that the choice of promoter definition drastically impacted the number of differentially methylated promoters found, with the wider definitions C, D and E resulting in several times the number of differentially methylated promoters than the narrower definitions A and B (Fig. 1D). The number of hypomethylated promoters identified was the most affected, likely reflecting the tendency of genomic regions further away from TSS to lose methylation in cancer. We found that among all promoters identified as differentially methylated with at least one of the five promoter definitions, only a tiny minority were common to all five definitions, with a large proportion exclusively identified with a single one of the definitions (Supplementary Fig. S7A and B). For example, 1292 promoters were found to be hypermethylated only when using promoter definition D, but not with any of the other definitions, while 2079 promoters were found to be hypomethylated only with promoter definition E, but not with any of the others.

Fixed promoter definitions display poor correlations between methylation and transcriptional activity

Promoter DNA methylation is generally of interest as it is assumed that it is associated with transcription of the associated gene. Given the substantial influence of choice of promoter definition on the calculation of promoter methylation levels, it seems obvious that promoter definition could also have a large influence on the calculation of correlation values between promoter methylation and transcript expression. To investigate this, we decided to examine how the correlation values between DNA methylation and transcriptional activity varied with distance of CpG sites from the TSS and if one promoter definition tended to capture the regions with the strongest correlations.

We discovered that there could be substantial variation in these correlation values, with groups of CpG sites with strongly negative or positive correlations found interspersed among CpGs with weak correlations and that fixed promoter definitions often failed to capture this complexity. Fig. 2B and C show the correlations values around *FOXD1* and *PACSIN3* in prostate metastasis samples as examples. Despite the

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

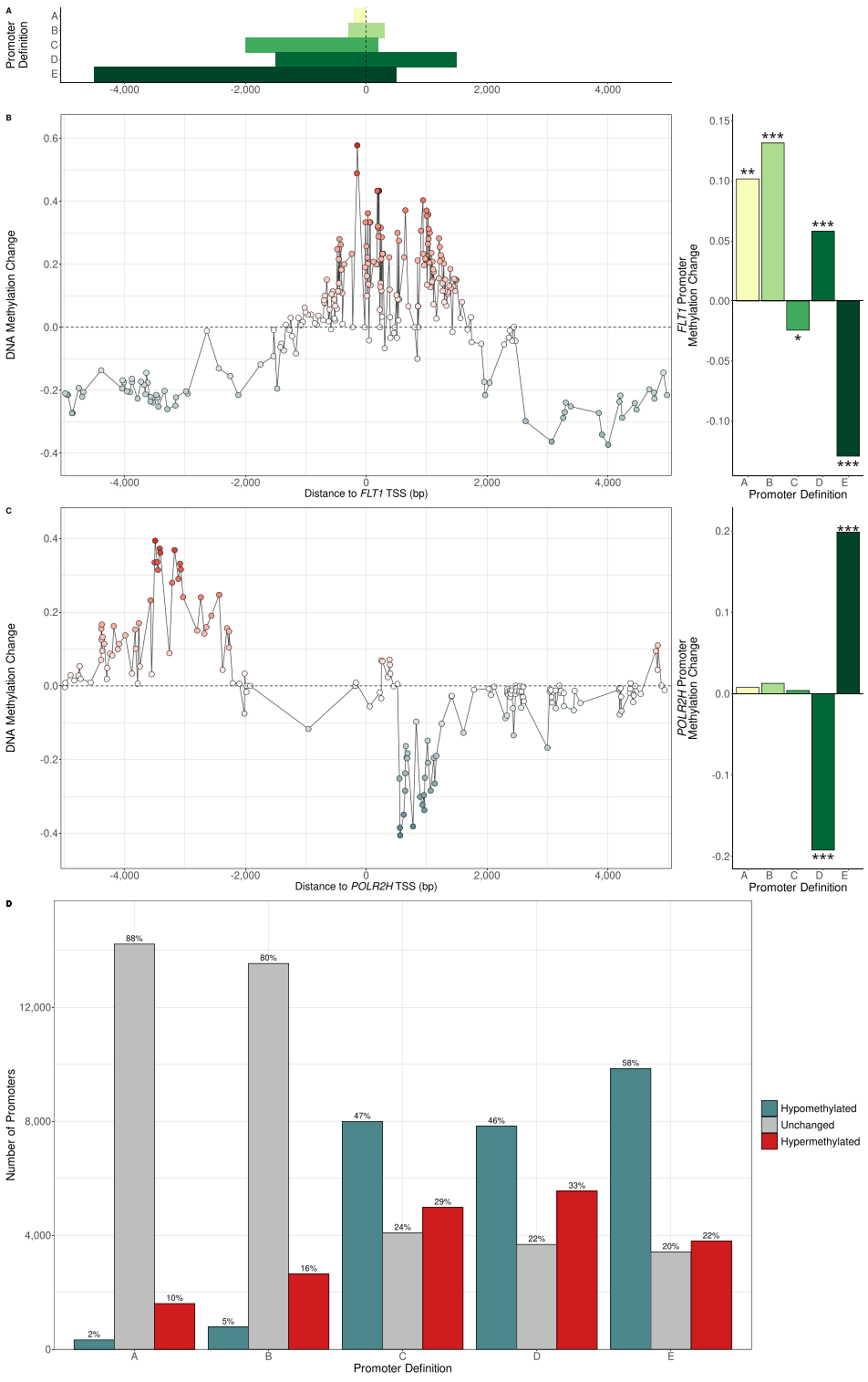


Figure 1. (**A**) The location of five different promoter definitions relative to the TSS from five different studies of DNA methylation. (**B** and **C**) The effect on the choice of promoter definition on differential promoter methylation analysis for *FLT1* (using the TSS associated with the canonical transcript ENST00000282397) and *POLR2H* (using the TSS associated with the canonical transcript ENST00000443489). In the left-hand plots, the *x*-axes show distance of CpG sites upstream and downstream of the TSS in base pairs and the *y*-axes show mean change of methylation in prostate tumour samples compared to normal prostate samples. The right-hand plots show the promoter methylation change calculated using different promoter definitions by averaging the methylation values across CpG sites overlapping each definition. Stars indicate the level of statistical significance for methylation change using a beta-binomial test (*** indicates a *P*-value < 0.001, ** indicates a *P*-value < 0.01 and * indicates a *P*-value < 0.05). With definitions A, B, and D, the *FLT1* promoter is hypermethylated, while using definitions C and E, it is hypomethylated. With definition D, the *POLR2H* promoter is hypomethylated, while for definition E, it is hypermethylated. (**D**) The number of significantly hypermethylated and hypomethylated promoters when using the different promoter definitions in the same prostate cancer dataset. Figures above the bars give percentages out of the total of all promoters tested. Promoter definition choice substantially affects the number of differentially methylated promoters discovered, with longer definitions resulting in greater numbers of differentially methylated promoters, especially the number of hypomethylated promoters.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

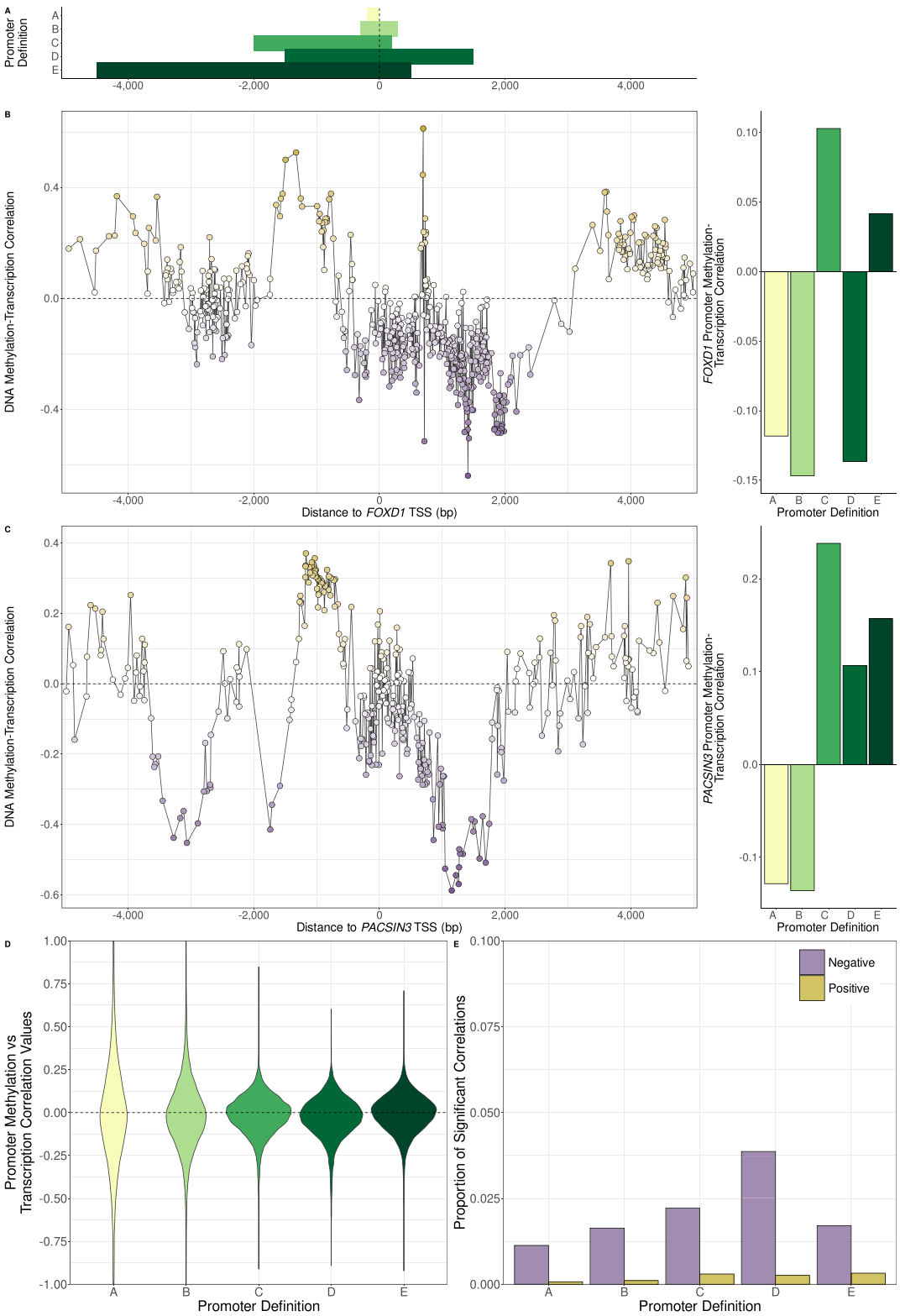


Figure 2. (**A**) The location of five different promoter definitions relative to the TSS from the same five different studies of DNA methylation as in Fig. 1A. (**B** and **C**) The effect of the choice of promoter definition on promoter methylation–transcription correlations for *FOXD1* (using the TSS associated with the ENST00000615637 transcript) and *PACSIN3* (using the TSS associated with the ENST00000298838 transcript) in prostate metastasis samples. In the left-hand plots, the *x*-axes show distance of CpG sites upstream and downstream of the associated TSS in base pairs and the *y*-axes show Spearman correlation values between CpG methylation and transcriptional activity of the TSS. The right-hand plots show the promoter DNA methylation-transcription Spearman correlation values were calculated for the different promoter definitions using the average methylation values across CpG sites overlapping each definition. Different promoter definitions can result in opposing promoter methylation–transcription correlation values, though no correlation values were statistically significant. (**D**) The distribution of promoter methylation-transcription Spearman correlation values for all protein-coding transcripts using each of the five different promoter definitions in normal prostate samples. Most correlation values are close to zero. (**E**) The proportion of statistically significant correlations in normal prostate samples for each promoter definition divided into negative and positive correlations. Only a small minority of correlations are significant using any promoter definition. Among the significant correlations are a small number of positive ones.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

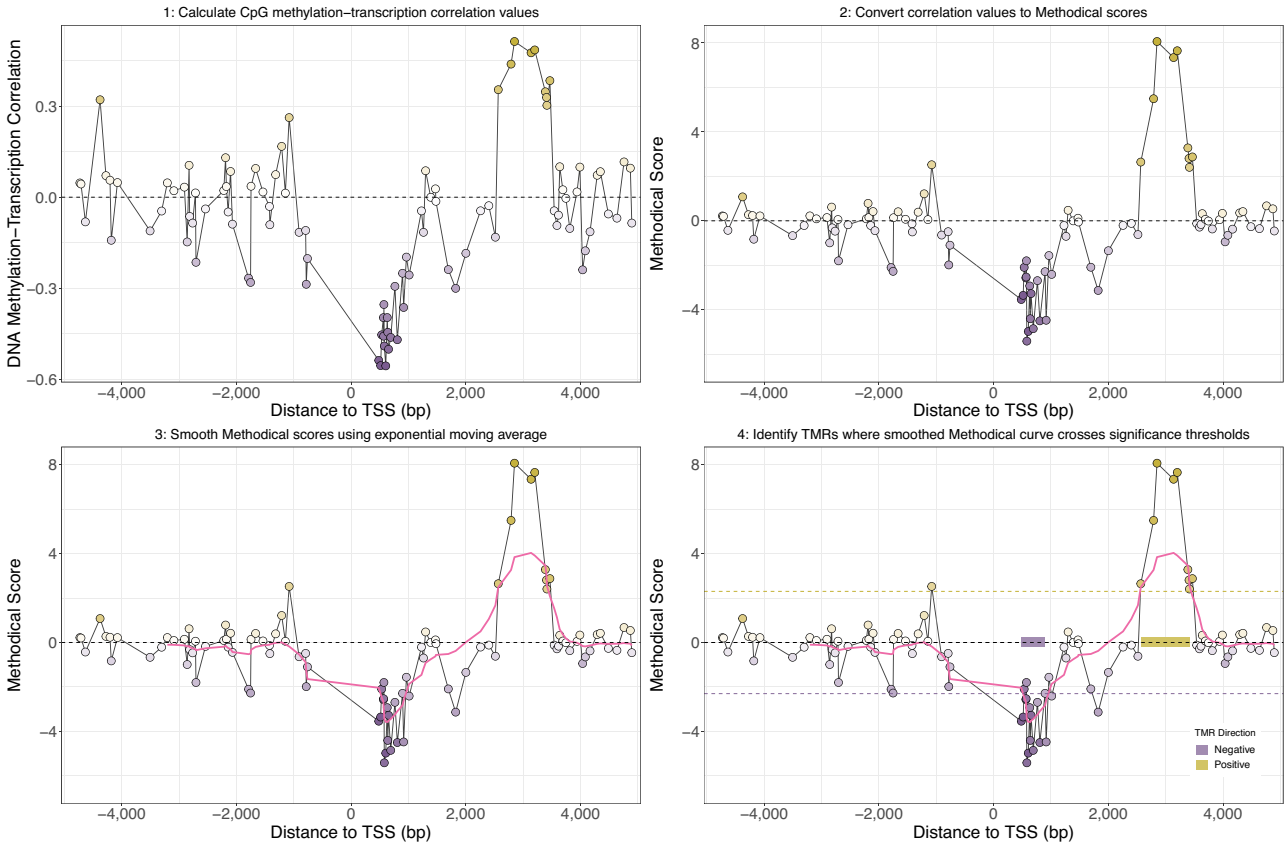


Figure 3. Overview of the identification of TMRs by Methodical using *QSOX1* (using the TSS associated with the transcript ENST00000367602) as an example. *X*-axes show distance from the *QSOX1* TSS in base pairs. *Y*-axis of the top left panel indicates the Spearman correlation between DNA methylation levels of CpG sites and expression of *QSOX1*, while *Y*-axes for other panel indicates Methodical scores associated with these correlations. (1) First Spearman correlation values are calculated between methylation of CpG sites close to a TSS and the expression of the transcript associated with that TSS. (2) Next, these correlation values are converted into Methodical scores by taking the logarithm base 10 of the *P*-values associated with the correlation values multiplied by −1 if the correlation is positive. (3) These methodical scores are then smoothed by using an exponential moving average, with the smoothed scores indicated by the pink line following the points. (4) Finally, two significance thresholds are used to identify TMRs, one for positive TMRs indicated by the upper dashed gold line and another for negative TMRs indicated by the lower dashed purple line. The location of the identified negative and positive TMRs are shown by the purple and gold blocks, respectively.

presence of CpG sites near the TSS where DNA methylation levels are relatively strongly correlated with transcription, fixed promoter definitions often resulted in poor correlations because they either failed to include these CpG sites, included many CpG sites where DNA methylation is not correlated with transcription or overlapped CpG sites with both negative and positive correlations with transcription.

Nevertheless, we wanted to determine if one definition generally tended to result in the strongest correlations between promoter DNA methylation and transcription. Thus, we calculated these correlations for protein-coding transcripts using each of the five chosen different definitions in the normal prostate samples (Fig. 2D and E), prostate tumour and metastasis samples (Supplementary Fig. S8). We found that, regardless of the promoter definition used, the correlation values were generally quite low, with over 40% of all correlation values having an absolute value less than 0.15 for all definitions across all datasets (Fig. 2D and Supplementary Figs S8A and S8B) and only a very small proportion of all correlation values being statistically significant (Fig. 2E and Supplementary Figs S8C and S8D). Promoter definition D, the definition with the greatest amount of sequence downstream of the TSS, resulted in the greatest number of significant correlation values in all groups of samples. This suggests that in addition to upstream regions, the region immediately downstream of the TSS could also be important to controlling transcriptional regulation via DNA methylation. Additionally, we observed that a small number of the statistically significant correlations are actually positive, supporting previous reports of positive association between DNA methylation and gene expression in certain settings [9, 10]. There were around twice as many significant correlation values overall in the prostate tumour and metastasis samples compared to normal prostate samples, possibly reflecting transcriptional dysregulation associated with aberrant DNA methylation in prostate cancer.

Identification of transcript-proximal methylation-associated regulatory sites

Due to the frequent failure of fixed promoter definitions to capture the regions where DNA methylation is most strongly correlated with transcription, we sought to develop an alternative approach that would systematically identify these regions. Thus, we developed Methodical, a computational approach that uses WGBS and RNA-seq to identify such regions (Fig. 3; details in "Materials and Methods"). We refer to these regions as transcript-proximal methylation-associated regulatory sites (TMRs). TMRs are classed as either having a negative or pos-

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

itive direction based on the sign of the correlation values between methylation of their CpG sites and transcription. Fig. 4A and B show TMRs identified for *FOXD1* and *PACSIN3* in prostate metastasis samples.

We first searched for TMRs within ± 5 kilobases (kb) of TSS to gauge the typical distance of TMRs from the TSS and noted that more TMRs were surprisingly found in the transcribed region downstream of the TSS, rather than in the upstream promoter region (Supplementary Fig. S9A). We then expanded the search region for TMRs to the whole transcribed region for each transcript in addition to 5 kb upstream of the TSS and 5 kb downstream of the TES in order to comprehensively identify the most common locations for TMRs. We observed that the region immediately downstream of the TSS had the highest density of negative TMRs, followed by the region immediately upstream of the TSS, with the density of TMRs steadily decreasing across the transcribed region towards the TES (Fig. 4C). In contrast, the density of positive TMRs remained relatively constant throughout the transcribed region. In the tumour and metastasis samples we saw a similar pattern, though with a higher proportion of positive TMRs which become as common as negative TMRs towards the TES (Supplementary Fig. S10). We also observed a similar distribution pattern when we examined the spatial distribution of TMRs we identified in a small number of diverse tissue samples from the Roadmap Epigenomics project (Supplementary Fig. S11), indicating that is a general pattern for TMRs across tissues.

The presence of several TSS in close proximity regulating different isoforms of the same gene could potentially complicate the interpretation of the spatial distribution patterns of TMRs. To determine if we saw the same general pattern when examining only one TSS per gene, we selected the TSS associated with the Matched Annotation from NCBI and EBI (MANE) transcripts, a set of ENSEMBL transcripts comprising a high confidence transcript for each gene with complete sequence identity with a Refseq transcript. When we examined the distributions of the subset of TMRs associated with these TSS, we observed an almost identical distribution pattern in normal prostate samples to those when considering TMRs associated with all TSS (Supplementary Fig. S12).

We found 7345 TMRs in normal prostate samples, 11 075 in prostate tumour samples and 4414 in prostate metastasis samples, with about 24% positive in the normal prostate samples and 40–50% positive in the prostate tumour and metastasis samples. These were associated respectively with 2153, 3341, and 1561 transcripts and 1867, 2810, and 1411 genes (Supplementary Fig. S9B).

We found that the prostate metastasis TMRs exhibited a much higher overlap with the TMRs found in prostate tumours than with those found in normal prostate samples than normal prostate. The transcripts and genes associated with TMRs displayed a similar pattern (Supplementary Fig. S9C). This indicates that the TMRs identified in prostate tumour and metastasis samples may be associated with transcriptional networks important to the development and progression of prostate cancer. Indeed, when we tested if the genes associated with TMRs in prostate tumour or metastases but not in normal samples were overrepresented for any MSigDB Hallmark pathways, we found that several cancer-associated signaling pathways were enriched, including 'TNFA signaling via NFKB', 'Hedgehog signaling', 'P53 Pathway', and 'IL2-STAT5 signaling' as well as other terms relevant to cancer development including 'epithelial-mesenchymal transition', 'hypoxia', and 'inflammatory response', strongly supporting a role for these genes in cancer development (one-sided Fisher's exact test FDR corrected P-value < 0.05).

We also noted that negative TMRs from one dataset generally had a higher overlap with the negative TMRs than with the positive TMRs from the other two datasets and vice versa for positive TMRs, supporting a consistent association between TMR methylation and transcription across datasets. To further investigate this, we then evaluated the TMR methylation–transcription correlations across the three datasets, comparing the correlations for TMRs identified within a particular dataset, which we refer to as internal TMRs for that dataset, with those identified in another dataset, which we refer to as external TMRs for that dataset. For example, the TMRs identified in normal prostate samples are internal to normal prostate samples and external to prostate tumours and metastases.

Unsurprisingly, the correlation values for internal TMRs were very strong and almost all statistically significant (Supplementary Fig. S9D and E). However, we also saw that a large proportion of correlation values involving external TMRs were strong with 45% statistically significant overall, a far higher proportion than with any fixed promoter definition (Fig. 2E). This proportion increased to 52% when considering the TMRs identified in prostate tumour samples in prostate metastasis samples or vice versa, further supporting that the TMRs identified in prostate tumours and metastases are regions important to prostate tumorigenesis.

Overlap with TMRs with regulatory features, chromatin states and transcriptional regulator binding sites

It has been reported that the relationship between DNA methylation and transcription varies with genomic context [5, 46] and also has opposing effects on the binding of different transcription factors [47]. We thus decided to investigate if TMRs were associated with certain regulatory elements, chromatin states and transcriptional regulator binding sites.

We calculated the enrichment of different classes of genomic regulatory elements among TMRs. We found that CpG islands, CpG shelves, and CpG shores as well as predicted promoter and enhancer regions all displayed significant enrichment for negative TMRs, with CTCF binding sites enriched among negative TMRs in prostate tumour and metastasis samples. Positive TMRs in normal prostate were enriched for predicted enhancer regions (Fig. 5A).

Following on from this, we evaluated in a similar manner the enrichment of TMRs for 18 chromatin states identified in healthy prostate tissue [48]. The most enriched chromatin states among TMRs were the TSS and enhancer-associated chromatin states. Interestingly, the bivalent/poised TSS state was enriched among negative TMRs identified in prostate tumour and metastasis samples (Fig. 5B).

We made similar observations when we examined the overlap of TMRs we identified in the samples from the Roadmap Epigenomics project with regulatory features and chromatin states, with TMRs again enriched for CpG islands, predicted promoter and enhancer regions and bivalent chromatin states (Supplementary Fig. S13A and B).

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

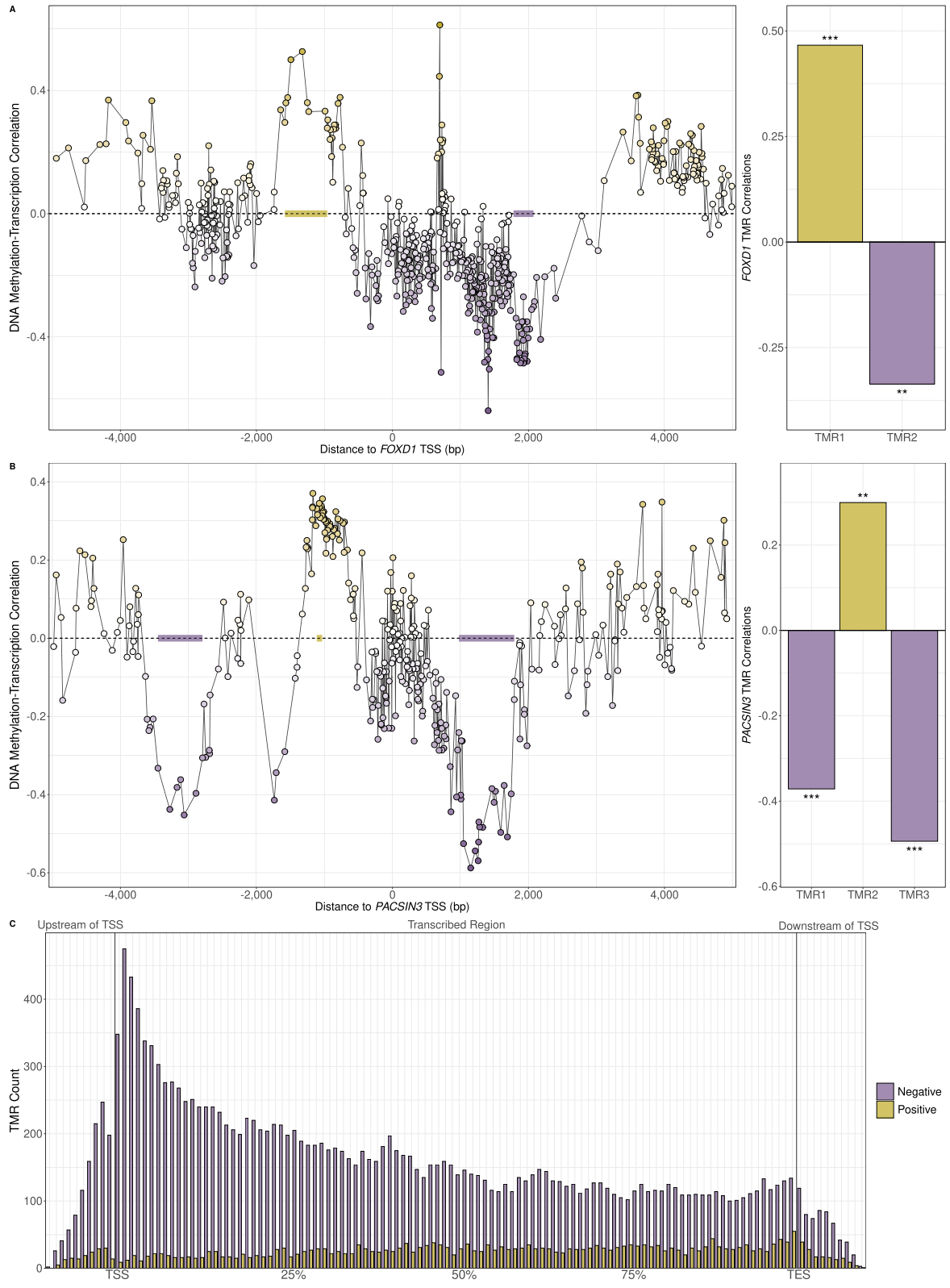


Figure 4. (**A** and **B**) The location of TMRs for *FOXD1* (associated with the ENST00000615637 transcript) and *PACSIN3* (using the ENST00000298838 transcript) found in prostate metastasis samples. *X*-axes in left-hand plots show distance of CpG sites from TSS in base pairs while *Y*-axes show Spearman correlation values between CpG methylation levels and transcription. The right-hand plots show the TMR DNA methylation-transcription Spearman correlation values calculated by averaging the methylation values across CpG sites overlapping each TMR. Stars indicate the level of statistical significance of correlations using a t-distribution. *FOXD1* has one negative and one positive TMR and *PACSIN3* has two negative and one positive TMR, all of which are significantly correlated with transcriptional activity. (**C**) Metagene plot showing the location of TMRs identified in normal prostate samples within transcribed regions or within 5 kb upstream of the TSS or 5 kb downstream of the TES for all transcripts. Each transcribed region was divided into 100 equally sized sections, while upstream and downstream regions were divided into 500 bp bins.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

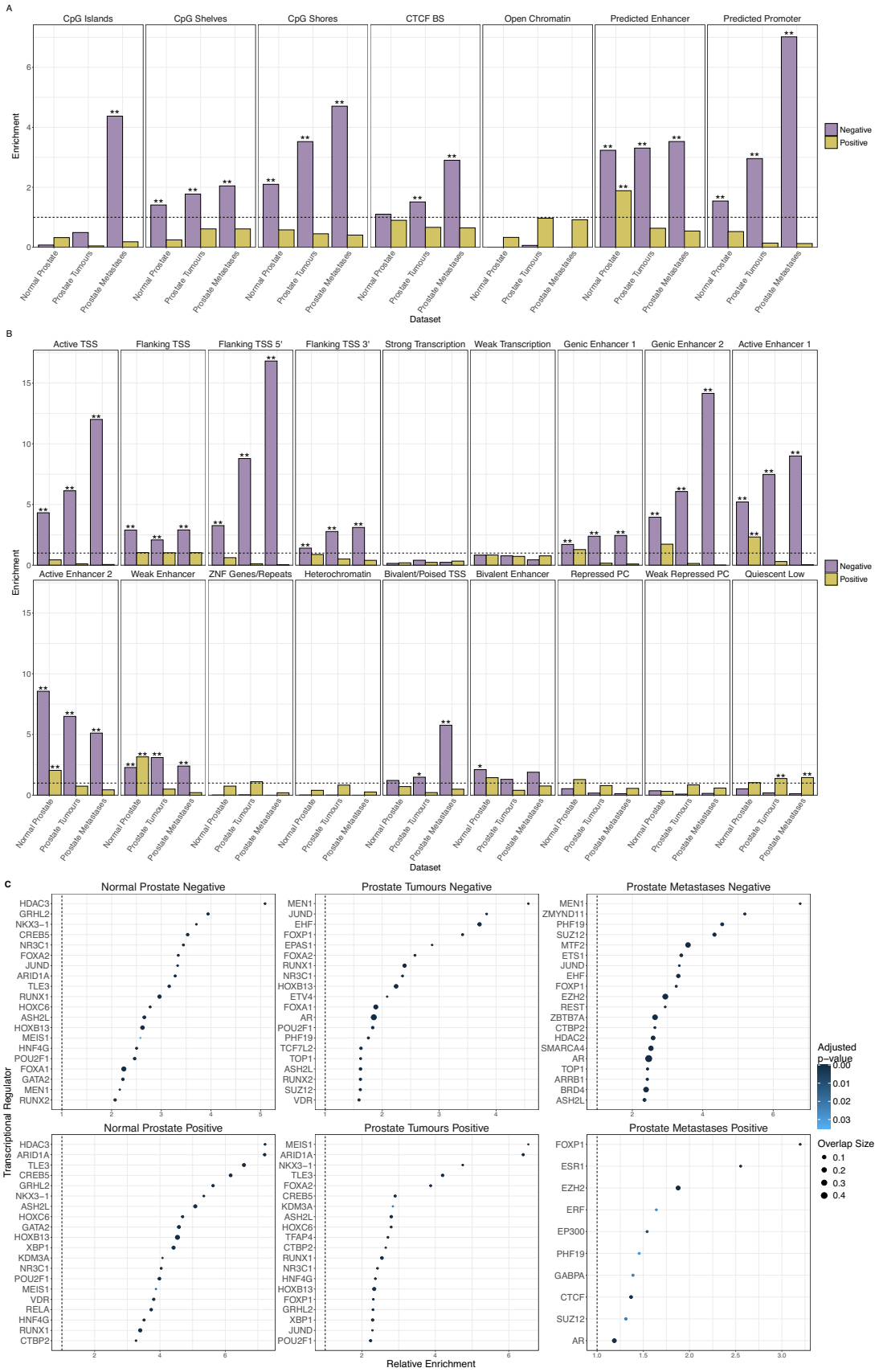


Figure 5. (**A**) The enrichment of different classes of genomic regulatory elements among TMRs. CTCF BS stands for CTCF binding site. (**B**) The enrichment of different chromatin states among TMRs, calculated similarly as with the genomic regulatory elements above. (**C**) Enrichment of binding sites for transcriptional regulators among different groups of TMRs. The top 20 most enriched regulators for each TMR group is shown. Supplementary Table S1 lists all transcriptional regulators with significantly enriched binding sites.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

Next we evaluated the overlaps of TMRs with binding sites in prostate tissue for 77 transcriptional regulators, including various transcription factors. MEN1, a known tumour suppressor gene associated with endocrine tumours [49], was one of the transcriptional regulators most strongly associated with negative TMRs, being the most enriched regulator for negative TMRs identified in both prostate tumour and prostate metastasis samples. Forkhead box (FOX) transcription factor family members FOXA2 and FOXP1 were also among the proteins with binding sites most strongly associated with negative TMRs and have both previously been implicated in prostate cancer [50, 51]. Notably, binding sites for several chromatin remodellers were enriched for different TMR groups, including HDAC1, HDAC2, HDAC3, ARID1A, EZH2, SUZ12, and TET2 (Fig. 5C; See Supplementary Table S1 for full results).

Methylation change at TMRs in tumorigenesis

DNA Methylation change affects much of the genome in cancer, with increases of methylation at specific loci, particularly CpG islands, contrasting with a general loss of methylation across the rest of the genome [52–54]. Given the vast number of regions affected by DNA methylation change, it has been difficult to determine which changes are playing an active role in tumour development and progression [55]. Considering this, we reasoned that examination of methylation change at TMRs in cancer could help identify regions where altered methylation is associated with transcriptional changes in cancer and thus more likely to be clinically relevant.

We thus tested differential methylation of TMRs between the prostate tumour and matching normal prostate samples and discovered that TMRs are very frequently affected by altered methylation in prostate cancer (Fig. 6A). While negative TMRs displayed an almost equal tendency to become either hypermethylated or hypomethylated overall in prostate tumours, positive TMRs exhibited a pronounced tendency to lose methylation, being the class of region with the greatest loss of methylation.

Differentially methylated regions (DMRs) were previously reported in the prostate metastasis samples [29] and so we were interested in the overlap between these regions and the TMRs we identified in those samples. The vast majority (96%) of the DMRs were hypomethylated, in contrast to the subset of DMRs which overlapped negative TMRs identified in prostate metastasis samples, where 22% were hypermethylated. We hypothesized that the increased methylation at TMRs is possibly more likely to be functionally relevant than the loss of methylation at the majority of hypomethylated DMRs. Supporting this, when we performed gene set overrepresentation analysis for the nearest genes to all DMRs within our TMR search space (transcribed regions ± 5 kb), we found no significantly enriched MSigDB Hallmark pathways. However, when we tested the genes associated with subset of DMRs that overlapped TMRs, we found several enriched hallmark pathways, including 'p53 pathway', 'androgen response', 'epithelial mesenchymal transition', 'KRAS signaling up', and 'Notch signaling' (one-sided Fisher's exact test, FDR-corrected P-value < 0.05).

We also wondered how often DMRs discovered near the vicinity of transcribed regions are associated with transcriptional activity. Thus, we selected for all DMRs which overlapped the regions in which we searched for TMRs (transcribed regions plus 5 kb upstream and downstream) and calculated the correlation between the methylation of these DMRs and expression of the associated transcript. We found that less than 10% of this subset of DMRs were significantly associated with expression. Thus, standard differential methylation testing can lead to the identification of a vast number of regions which are generally not correlated with the expression of nearby genes. Consequently, methylation change at TMRs may be more functionally relevant than at DMRs in general, supported by the overrepresentation of genes in cancer-related KEGG and Hallmark pathways among TMRs overlapping DMRs.

We subsequently performed pathway overrepresentation analysis for the genes associated with differentially methylated TMRs using KEGG pathways. Strongly supporting a role for altered TMR methylation in tumorigenesis, we found that the term "pathways in cancer" was enriched among the genes associated with negative TMRs identified in normal prostate samples and prostate tumours which were hypermethylated in prostate cancer, with several other cancer-related terms also enriched, including 'prostate cancer', 'pancreatic cancer', 'melanoma', 'glioma', and 'small cell lung cancer' (Supplementary Fig. S14A).

Accordingly, we found several examples of TMRs associated with prostate cancer-related genes displaying methylation changes in prostate tumours. This included hypermethylation of negative TMRs associated with tumour suppressor genes, such as *TGFBR2* and *SMAD3*, as well as loss of methylation at negative TMRs for the oncogenes *ERBB3* and *SND1* and for *PTPN13*, which is described as having both oncogenic and tumour suppressor roles [56]. Additionally, *GSTP1*, one of the most frequently hypermethylated genes in prostate cancer and a proposed caretaker gene [57], displayed pronounced hypermethylation of a TMR located just downstream of its TSS. Intriguingly, all these TMRs were located downstream of the TSS (Supplementary Fig. S15).

Several other KEGG pathways related to the extracellular matrix, such as 'focal adhesion', 'regulation of actin cytoskeleton', and 'ECM receptor interaction' were also enriched, suggesting that differential methylation of TMRs could be associated with tumour invasiveness.

Conversely, genes associated with hypomethylated negative TMRs were enriched for the androgen response pathway and related pathways such as cholesterol homeostasis and estrogen response, indicating that loss of methylation at these TMRs is associated with increased androgen signalling (Supplementary Fig. S14B). Only a few pathways were significantly enriched among genes associated with positive TMRs, probably related to the identification of much fewer positive TMRs than negative TMRs, but included terms related to angiogenesis, KRAS signalling and also EMT.

To investigate if methylation change at TMRs may be a common occurrence across different cancers, we evaluated methylation change at the TMRs identified in the prostate metastasis samples in tumour and matching normal samples with WGBS data from eight different patients with different cancer types from TCGA [58]. We found that the negative TMRs once again often gained methylation, while the positive TMRs were generally the class of region most affected by methylation loss (Fig. 6C).

Finally, we wanted to confirm if the methylation of the TMRs we identified in prostate may display the expected associations with gene expression in different cancer types and so we used methylation and RNA-seq data from TCGA.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

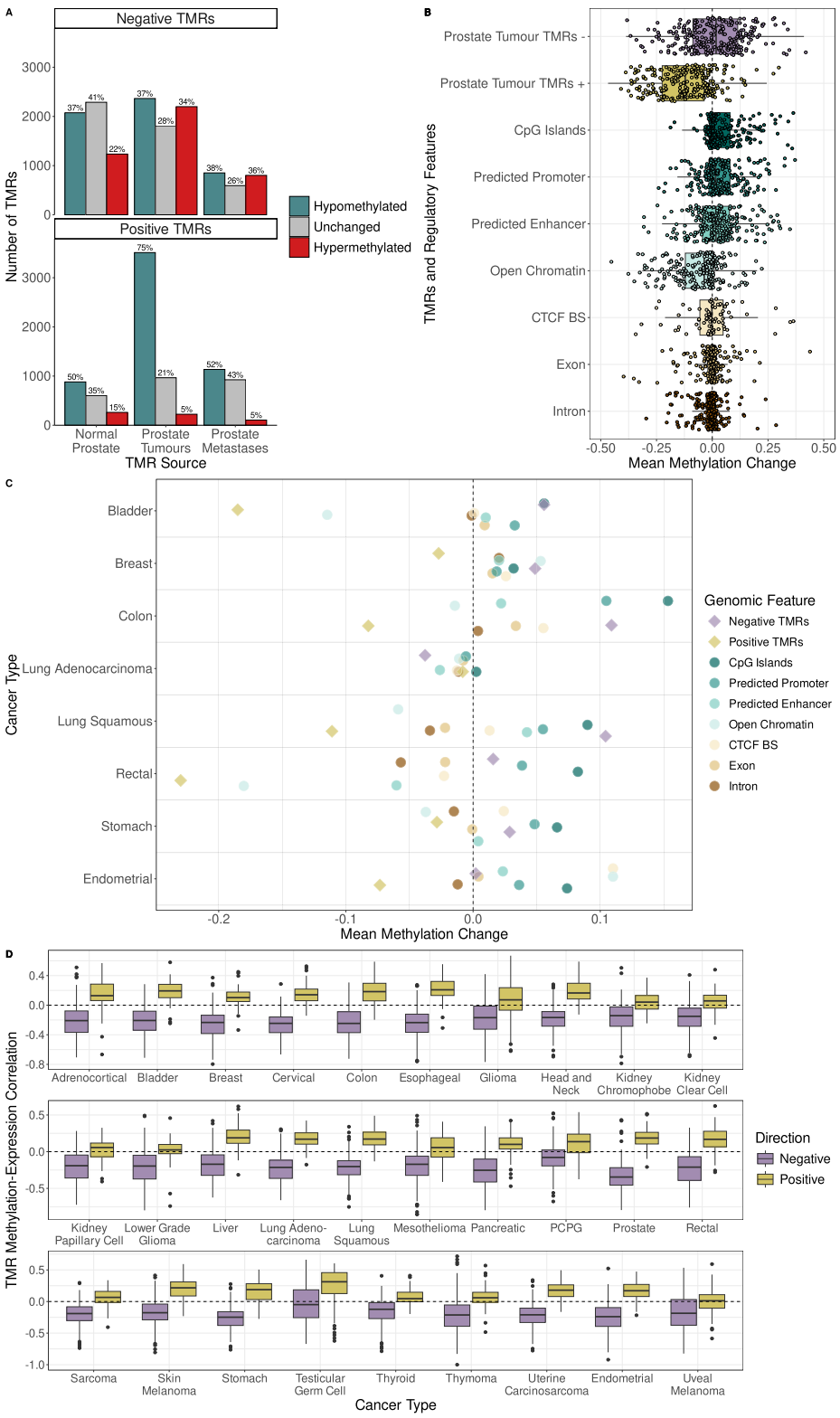


Figure 6. (**A**) The number of differentially methylated negative and positive TMRs in prostate cancer. The *x*-axis shows the source of TMRs, the *y*-axis the number of TMRs and colour indicates the direction of methylation change. Percentages above bars indicate the percentage of all TMRs in the relevant group affected by the indicated methylation change. (**B**) Methylation change at TMRs in prostate tumours compared to different regulatory regions. Boxplots show the distributions of the mean methylation change of regions belonging to the indicated class. The mean methylation change of 250 random regions belonging to each class are plotted as points on top of the boxplots. (**C**) Methylation change at various classes of genomic regions in eight cases of different cancer types with WGBS data from TCGA. TMRs are those identified in prostate metastasis samples and are indicated by diamonds with other classes of regions represented by circles. (**D**) Distribution of Spearman correlation values between methylation of TMRs associated with MANE transcripts identified in prostate tumour samples and expression of the relevant gene across different cancer types from TCGA. Direction of correlation values are generally as would be expected, with negative correlations for negative TMRs and positive correlations for positive TMRs, though the strongest correlations were generally observed in prostate cancer samples from TCGA. PCPG stands for pheochromocytoma and paraganglioma.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

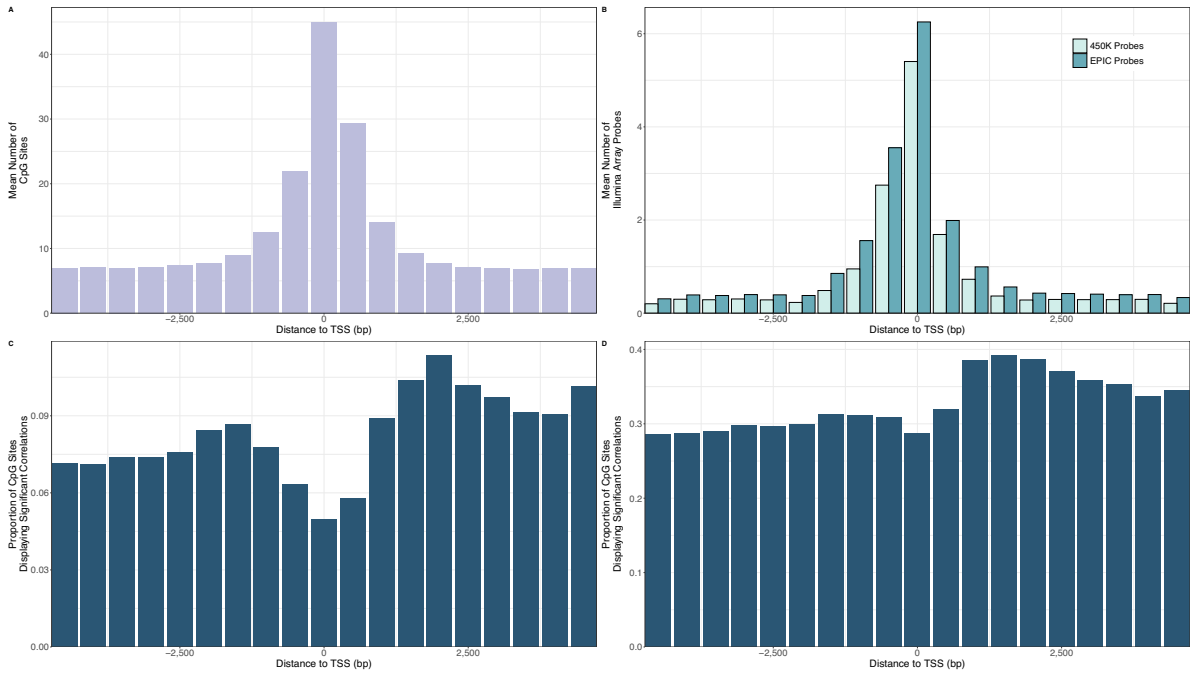


Figure 7. (**A**) The mean number of CpG sites in 500 bp bins around TSS. The bin overlapping the TSS is the most common location for CpG sites, with the bin just downstream the second most common. (**B**) The mean number of CpG sites targeted by the Illumina HumanMethylation450 and EPIC methylation microarrays in 500 bp bins around TSS. Most CpGs targeted are in the bin overlapping the TSS or the bin 500 bp upstream, with much fewer CpGs targeted elsewhere, including the region just downstream of the TSS with a high concentration of CpG sites. (C and D) The proportion of CpG sites where DNA methylation was significantly correlated with transcript expression in normal samples (**C**) and tumour samples (**D**) from TCGA, testing significance separately in each tissue/cancer type. Significant was defined as a corrected *P*-value less than 0.05 for the Spearman correlation values. The greatest proportion of significant correlation is found downstream of the TSS.

However, only gene-level expression data, but not transcript-level, is available publicly from TCGA, though TMRs are specifically associated with the expression of individual transcripts. Thus, we decided to evaluate only TMRs associated with MANE transcripts. The MANE transcripts are generally the most expressed transcripts for each gene, and so we reasoned that the the methylation of TMRs associated with MANE transcripts should generally also be strongly associated with the overall expression for the corresponding gene.

Additionally, the vast majority of the DNA methylation data for TCGA samples was produced using the Illumina 450K methylation array which covers only a minority of CpG sites in the genome. Thus, we evaluated the correlation values between the methylation of TMRs associated with MANE transcripts identified in prostate metastasis samples which also overlapped CpG sites covered by the Illumina 450K methylation array and expression of the relevant gene in 29 different cancer types from TCGA. Only 1315 TMRs (974 negative and 341 positive) were associated with MANE transcripts and were targeted by one or more probes from the array. Additionally, methylation data was generally only available for a proportion of the CpG sites in each of these TMRs. Regardless, we observed that the majority of calculated correlation values were significant (62%) and that the signs of the correlation values generally matched those expected given the direction of the TMRs (Fig. 6D). Thus, many of the TMRs we identified in prostate cancer are regions where methylation is associated with gene expression in a pan-cancer manner. We did observe that the strongest and greatest proportion of significant correlations (92%) were in the prostate cancer samples from TCGA, indicating that there is some tissue-specificity to the associations. Therefore, applying Methodical to datasets with WGBS and RNA-seq for other tissues could lead to the identification of TMRs which are more relevant to different tissue or cancer types.

Methylation microarrays often miss regions where DNA methylation is most strongly associated with transcription

Several previous studies of the association between DNA methylation and transcription have used methylation data produced by Illumina arrays [10, 27, 28] and so we wanted to evaluate if analysis of the CpG sites targeted by these arrays could adequately capture the general correlation patterns between DNA methylation and transcription around TSS. We first noted that the CpG sites covered by Illumina arrays were biased towards those within about 1 kb upstream of the TSS, with relatively low coverage of CpGs downstream of the TSS where we identified the greatest number of TMRs (Fig. 7A and B).

When we then examined the distribution of CpGs where DNA methylation was significantly correlated with transcription in both normal and tumour samples from TCGA, we found that the region with the greatest proportion of CpGs with significant correlations was the region downstream of the TSS (Fig. 7C and D), fitting with our discovery of a high number of TMRs in this region. We noted similar patterns with normal and tumour samples when we looked at individual tissue types from TCGA (Supplementary Fig. S16), demonstrating this is a general pattern across tissue and cancer types.

Thus, the bias of methylation arrays to the region immediately surrounding the TSS and just upstream has likely resulted in an underappreciation of the association of DNA

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

methylation further from the TSS with transcription, particularly downstream methylation. Indeed, we found that, depending on the dataset, 65–75% of TMRs we identified did not contain a single CpG site that was targeted by the Illumina HumanMethylation450 array.

Discussion

Here, we have demonstrated firstly how the use of different arbitrary promoter definitions in studies of DNA methylation can lead to profoundly discordant results. These different promoter definitions can result in vastly different numbers of differentially methylated promoters being identified and perhaps even more strikingly, can lead to completely opposing results even at the same TSS. We found that these contradictory results with different promoter definitions can affect thousands of different TSS. In light of this, the wide variation in how promoters are defined between different studies of DNA methylation is concerning.

Unsurprisingly given the effect of promoter definition choice on the calculation of promoter methylation levels, we also found that the correlation values between promoter methylation levels and transcript expression are hugely influenced by the choice of promoter definition. The vast majority of these correlation values are very weak and statistically insignificant, consistent with other studies modelling gene expression using DNA methylation [7, 59–62], and we have revealed that this is in part because fixed-size promoter definitions often miss the complex relationship between methylation of CpG sites, distance from TSS and transcriptional activity. Regions where DNA methylation is negatively correlated, positively correlated or uncorrelated with transcription can all occur in the region immediately upstream and downstream of the TSS, with different TSS displaying completely different patterns.

This is in line with previous work extracting higher order features from methylation data, which enabled the identification of five major spatially correlated promoter methylation patterns [14]. Unsurprisingly, a uniformly hypermethylated pattern was associated with transcriptional repression. However, a uniformly lowly-methylated pattern was also unexpectedly associated with repression. Conversely, a U-shaped pattern characterized by a central hypomethylated region flanked by hypermethylated regions on either side was associated with high expression and interestingly this pattern was largely found at CpG islands and often associated with housekeeping genes. A forward S-shaped pattern characterized by lowly methylated region which gradually transitions into a highly methylated region and the reverse S-shaped pattern were associated with an intermediate level of gene expression. Taken together, these patterns underscore that the effect of DNA methylation on transcriptional activity is highly context dependent.

Approaches using more sophisticated machine learning algorithms, such as support vector machines or deep learning approaches, with higher resolution methylation features have proven to predict transcriptional activity much more accurately than simpler linear approaches [14, 17, 18]. The superiority of these complex models likely reflects the intricate relationship between methylation at different CpG sites in both promoter-proximal and distal enhancer regions and transcriptional activity. However, difficulty in interpreting such models makes determining the precise regions where DNA methylation is the most important for predicting transcriptional activity challenging. In contrast, our study specifically focuses on identifying and characterizing regions with the strongest correlations between DNA methylation and transcription.

In response to the above issues, we believed that WGBS and RNA-seq data could be combined to enable the systematic identification of precise regions where DNA methylation is associated with transcription, providing an alternative to the use of arbitrary fixed-size promoter definitions. Hence we developed Methodical, a novel algorithm that integrates WGBS and RNA-seq data to identify regions where CpG methylation displays consistent correlation with transcriptional activity of associated TSS. We call these regions transcript-proximal methylation-associated regulatory sites sites (TMRs), with TMRs having either a negative or positive direction depending on the correlation between DNA methylation and transcription. We applied Methodical to three datasets, one for normal prostate, one for prostate tumours and another for prostate metastases, with a large number of samples with both WGBS and RNA-seq data to identify a set of TMRs for each dataset.

As should be expected, the methylation of TMRs displays a much stronger correlation with transcriptional activity than the methylation of fixed promoter definitions when evaluating the correlations within the dataset in which the TMRs were identified. However, the association of TMR methylation with transcriptional expression could often be validated in other datasets, including those for other tissue types. We did note that the TMRs we identified in prostate tumours displayed stronger correlations when evaluated in prostate cancer than in other cancer types from TCGA, indicating that TMRs are partially tissue specific.

We discovered that the most common location for TMRs was the region immediately downstream of the TSS. Supporting this, we also found that the region within the first 2 KB downstream of the TSS was generally the region with the greatest proportion of CpG sites where DNA methylation was significantly correlated with transcription. We observed this pattern in diverse tissue and cancer types using both WGBS and methylation array data. This is in line with other recent studies modelling transcription using DNA methylation which found that methylation of the region downstream of the TSS was the most important for predicting gene expression [15, 16].

Around ~25–50% of TMRs were positive, with more positive TMRs found in prostate tumour and metastasis samples, demonstrating that the conventional wisdom that DNA methylation is negatively associated with transcription does not always hold true and supporting previous reports of positive associations between DNA methylation and transcription [9, 10]. Several possible mechanisms for transcriptional activation by DNA methylation have been proposed, including blocking the binding of transcriptional repressors and activation of alternative promoters [9].

We found that TMRs are enriched for certain genomic elements and chromatin states inferred from healthy prostate samples. Notably, negative TMRs from prostate tumours and metastases were highly enriched for a bivalent/poised TSS chromatin state. This bivalent state is characterized by the co-occurrence of both activating and repressive histone modifications, such as H3K4me3 and H3K27me3, respectively, and was first identified at the promoters of developmentally regulated genes in embryonic stem cells (ESCs). This led to the hypothesis that bivalency poises these developmental genes

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

for rapid activation during development while maintaining a transcriptionally inactive state in the absence of activating signals [63, 64].

However this hypothesis has been challenged by the work of Kumar *et al.* [65]. They demonstrated that bivalency does not seem to poise genes for rapid activation as the activation of bivalent genes was neither stronger nor more rapid than that of transcriptionally silenced genes lacking H3K4me3 at their promoters. Instead, they proposed that H3K4m3e may represent a general mechanism to maintain the unmethylated state of CGIs, even those associated with transcriptionally inactive genes.

While bivalent promoters generally display low methylation levels in normal cells [66, 67], it has been repeatedly observed that most genes that become hypermethylated in cancer have CGI promoters which are bivalently marked in ESCs [68–71]. This suggests that bivalency predisposes these genes for methylation gain in cancer. Additionally, bivalent regions often display the strongest hypermethylation of all chromatin states [71, 72]. However, as loss of bivalent chromatin seems to be a general feature of cancer, it may be that it is actually the loss of of bivalency, and H3K4me3 in particular, that leads to increased methylation, rather than the presence of bivalency *per se* [65, 71]. This was supported by the finding that regions that lost bivalency in cancer cell lines displayed greater hypermethylation than regions which remained bivalent [71]. Thus, the prevalence of TMRs in bivalent regions may indicates that the gain in methylation in these regions is often associated with transcriptional repression of the associated genes.

Given the vast numbers of regions affected by DNA methylation change in cancer, we thought that focusing on methylation change at TMRs in cancer could improve our understanding of how DNA methylation alterations lead to transcriptional misregulation during tumorigenesis. We discovered that TMRs displayed a pronounced tendency to become differentially methylated in prostate cancer. Positive TMRs were affected by greater methylation loss on average than all other classes of genomic elements studied. TMRs also displayed altered methylation in other tumour types. This is fitting with a prior study reporting that regions displaying variable methylation between different tissue types and during reprogramming of induced pluripotent stem cells, and thus presumably associated with developmental gene expression networks, are often differentially methylated in a variety of cancer types [59].

Pathway overrepresentation analysis for genes associated with hypermethylated negative TMRs in prostate cancer revealed many pathways related to cancer, invasiveness or oncogenic signalling pathways, supporting a role for TMR hypermethylation in tumorigenesis. Conversely, hypomethylated negative TMRs were enriched for genes involved in the androgen response, fitting with previous observations of hypomethylation at genes involved in the androgen-response in prostate cancer [21, 29].

To our knowledge, this is one of the most detailed studies to date of the relationship between DNA methylation around TSS and transcriptional activity. Several other studies of the association between DNA methylation and transcription have used methylation data coming from Illumina methylation microarrays [10, 27, 28]. These microarrays profile only a small fraction of CpG sites in the human genome and are highly biased towards certain genomic contexts, particularly upstream of the TSS. We have shown that regions downstream of the TSS actually have the greatest proportion of CpG sites where DNA methylation is significantly correlated with transcription. Thus, studies based on these arrays have missed many regions where DNA methylation is associated with transcription and likely have led to an underappreciation of the association of DNA methylation downstream of TSS with transcriptional activity. Future studies of DNA methylation should therefore aim for greater coverage of regions downstream of TSS.

Application of Methodical to other datasets with large numbers of samples with base-resolution methylation data and RNA-seq data for different tissue or tumour types when they become available would enable the identification of TMRs which may be more relevant for different tissue and cancer types. Here we only examined regions proximal to TSS, however Methodical could also be used to identify distal regions, such as enhancers, where DNA methylation is associated with transcription. In that case, care should be taken to account for correlations arising from gene co-expression, such as through a permutation-based approach [73]. Furthermore, as Methodical can only identify correlative relationships, experimental studies would be needed to demonstrate that TMR methylation is causally involved in influencing transcription. For example, investigating if treatment with a demethylating agent such as 5-azacytidine or experimental alteration of TMR methylation using CRISPR-Cas9 technology leads to the expected expression changes could confirm if TMR methylation is causally involved in gene regulation.

In summary, Methodical enables the identification of regions where DNA methylation is correlated with transcriptional activity, enabling insights into the fundamental relationship between DNA methylation and gene expression. Better understanding of this relationship will in turn improve our understanding of how altered DNA methylation in cancer is associated with perturbation of the transcriptome during tumour development.

Acknowledgements

We would also like to thank Dr. David Quigley and Dr. Jing Li for providing much of the data used in this study. The graphical abstract was created using BioRender.com.

Author contributions: Richard Heery (Conceptualization [equal], Investigation [Lead], Writing – original draft [equal], Visualization [Lead]), Martin Schaefer (Conceptualization [equal], Writing – original draft [equal]), Funding acquisition [Lead], Supervision [Lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

The work leading to this manuscript was supported by Fondazione AIRC, grant reference numbers MFAG n. 21791, Bridge Grant n. 29162 and Investigator Grant n. 30887. It has been partially supported by the Italian Ministry of Health with Ricerca Corrente and 5 × 1000 funds.

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

Data availiability

Processed WGBS data for normal prostate and prostate tumour samples from the CPGEA project [12] was downloaded from download.big.ac.cn:/gsa-human/HRA000099/processed/all_processed_files.tar.gz while processed WGBS data for prostate metastasis samples from the MCRPC project [29] was provided by the authors. Metastasis sample bedGraph files for WGBS analysis of 39 tumours and 8 matching normal samples from TCGA were downloaded from https://zwdzwd.s3.amazonaws.com/directory_listing/trackHubs_TCGA_WGBS_hg38.html?prefix=trackHubs/TCGA_WGBS/hg38/bed/. Illumina HumanMethylation450 files with probe beta values for tumour and normal samples for cancer types from TCGA were downloaded from the Genomic Data Commons (GDC) (https://portal.gdc.cancer.gov/) using the GDC-client tool. Genomic locations of probes for Illumina methylation microarrays were downloaded from the Illumina website. BED files for WGBS data for 38 different non-cancer human tissue samples belonging to 19 different tissue types from the Roadmap Epigenomics project [65] were downloaded from the ENCODE data portal [73]. See Supplementary Table S2 for files and tissue types.

CpG methylation values and RNA-seq transcript counts for the CPGEA and MCRPC projects are available as a Bioconductor ExperimentHub package at https://bioconductor.org/packages/release/data/experiment/html/TumourMethData.html.

FASTQ files for RNA-Seq data from the CPGEA were downloaded from the Genome Sequence Archive for Human (http://bigd.big.ac.cn/gsa-human/). BAM files for RNA-seq data for the MCRPC project were downloaded from GDC using GDC-client (https://portal.gdc.cancer.gov/) and FASTQ files were generated from them using Samtools (version 1.11). STAR count files with gene expression counts for TCGA were downloaded from GDC using GDC-client for each cancer type. RNA-seq gene counts for Roadmap Epigenomics project samples were downloaded from the ENCODE data portal. See Supplementary Table S2 for files and tissue types.

Methodical is available as an R/Bioconductor package at https://bioconductor.org/packages/methodical

Supplementary Table S1 lists all significantly enriched transcriptional regulators for TMR groups. Supplementary Table S2 gives the tissue types and file accession ID for all samples from the Roadmap Epigenomics project used in the study.

BED files with coordinates of TMRs for the hg38 genome build are available to download from https://github.com/richardheery/heery_2025_scripts/tree/master/finding_tmrs/tmr_bed_files

Scripts to reproduce analyses and figures are available at https://github.com/richardheery/heery_2025_scripts

R scripts have also been deposited to Zenodo at the DOI: 10.5281/zenodo.16964547 and the data repository DOI: 10.5281/zenodo.16964637.

References

1. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. *Science* 2001;**293**:1068–70. https://doi.org/10.1126/science.1063852
2. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;**14**:204–20. https://doi.org/10.1038/nrg3354
3. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;**20**:590–607. https://doi.org/10.1038/s41580-019-0159-6
4. Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 2006;**31**:89–97. https://doi.org/10.1016/j.tibs.2005.12.008
5. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**:484–92. https://doi.org/10.1038/nrg3230
6. Booth MJ, Branco MR, Ficz G *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 2012;**336**:934–7. https://doi.org/10.1126/science.1220671
7. Wagner JR, Busche S, Ge B *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* 2014;**15**:R37. https://doi.org/10.1186/gb-2014-15-2-r37
8. Farlik M, Halbritter F, Müller F *et al.* DNA methylation dynamics of human hematopoietic stem cell differentiation. *Cell Stem Cell* 2016;**19**:808–22. https://doi.org/10.1016/j.stem.2016.10.019
9. Smith J, Sen S, Weeks RJ *et al.* Promoter DNA hypermethylation and paradoxical gene activation. *Trends Cancer* 2020;**6**:392–406. https://doi.org/10.1016/j.trecan.2020.02.007
10. Spainhour JC, Lim HS, Yi SV *et al.* Correlation patterns between DNA methylation and gene expression in the cancer genome atlas. *Cancer Inform* 2019;**18**:1176935119828776.https://doi.org/10.1177/1176935119828776
11. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics Chromatin* 2018;**11**:37.
12. Brenet F, Moh M, Funk P *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* 2011;**6**:e14524. https://doi.org/10.1371/journal.pone.0014524
13. Kremer LP, Braun MM, Ovchinnikova S *et al.* Analyzing single-cell bisulfite sequencing data with MethSCAn. *Nat Methods* 2024;**21**:1616–23. https://doi.org/10.1038/s41592-024-02347-x
14. Kapourani C-A, Sanguinetti G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* 2016;**32**:i405–12. https://doi.org/10.1093/bioinformatics/btw432
15. Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Res* 2017;**45**:5100–11. https://doi.org/10.1093/nar/gkx078
16. Uzun Y, Wu H, Tan K. Predictive modeling of single-cell DNA methylome data enhances integration with transcriptome data. *Genome Res* 2021;**31**:101–9. https://doi.org/10.1101/gr.267047.120
17. Gao S, Zhu H, Cai K *et al.* TRAmHap: accurate prediction of transcriptional activity from DNA methylation haplotypes in bisulfite-sequencing data. *Briefings Bioinf* 2023;**24**:bbad214. https://doi.org/10.1093/bib/bbad214
18. Williams J, Xu B, Putnam D *et al.* MethylationToActivity: a deep-learning framework that reveals promoter activity landscapes from DNA methylomes in individual tumors. *Genome Biol* 2021;**22**:24. https://doi.org/10.1186/s13059-020-02220-y
19. Chen Y, Breeze CE, Zhen S *et al.* Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. *Epigenetics Chromatin* 2016;**9**:10.
20. Saghafinia S, Mina M, Riggi N *et al.* Pan-cancer landscape of aberrant DNA methylation across Human tumors. *Cell Rep* 2018;**25**:1066–80. https://doi.org/10.1016/j.celrep.2018.09.082
21. Li J, Xu C, Lee HJ *et al.* A genomic and epigenomic atlas of prostate cancer in Asian populations. *Nature* 2020;**580**:93–9. https://doi.org/10.1038/s41586-020-2135-x
22. Noushmehr H, Weisenberger DJ, Diefes K *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 2010;**17**:510–22. https://doi.org/10.1016/j.ccr.2010.03.017

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

23. Cao W, Lee H, Wu W *et al.* Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun* 2020;**11**:3675. https://doi.org/10.1038/s41467-020-17227-z
24. Weinstein JN, Collisson EA, Mills GB *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764
25. Gadd S, Huff V, Walz AL *et al.* A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat Genet* 2017;**49**:1487–94. https://doi.org/10.1038/ng.3940
26. Pidsley R, Zotenko E, Peters TJ *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;**17**:208. https://doi.org/10.1186/s13059-016-1066-1
27. Silva TC, Young JI, Martin ER *et al.* MethReg: estimating the regulatory potential of DNA methylation in gene transcription. *Nucleic Acids Res* 2022;**50**:e51. https://doi.org/10.1093/nar/gkac030
28. Sakellaropoulos T, Do C, Jiang G *et al.* MethNet: a robust approach to identify regulatory hubs and their distal targets from cancer data. *Nat Commun* 2024;**15**:6027. https://doi.org/10.1038/s41467-024-50380-3
29. Zhao SG, Chen WS, Li H *et al.* The DNA methylation landscape of advanced prostate cancer. *Nat Genet* 2020;**52**:778–89. https://doi.org/10.1038/s41588-020-0648-8
30. Shiraki T, Kondo S, Katayama S *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* 2003;**100**:15776–81. https://doi.org/10.1073/pnas.2136655100
31. Forrest ARR, Kawaji H, Rehli M *et al.* A promoter-level mammalian expression atlas. *Nature* 2014;**507**:462–70.
32. the FANTOM consortium, Lizio M, Harshbarger J, Shimoji H *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* 2015;**16**:22. https://doi.org/10.1186/s13059-014-0560-6
33. Bray NL, Pimentel H, Melsted P *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7. https://doi.org/10.1038/nbt.3519
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
35. Park Y, Figueroa ME, Rozek LS *et al.* MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* 2014;**30**:2414–22. https://doi.org/10.1093/bioinformatics/btu339
36. Liberzon A, Subramanian A, Pinchback R *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40. https://doi.org/10.1093/bioinformatics/btr260
37. Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;**1**:417–25. https://doi.org/10.1016/j.cels.2015.12.004
38. Cavalcante RG, Sartor MA. Annotatr: genomic regions in context. *Bioinformatics* 2017;**33**:2381–3. https://doi.org/10.1093/bioinformatics/btx183
39. Durinck S, Spellman PT, Birney E *et al.* Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91. https://doi.org/10.1038/nprot.2009.97
40. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 2009;**25**:1841–2. https://doi.org/10.1093/bioinformatics/btp328
41. Gel B, Díez-Villanueva A, Serra E *et al.* regioneR: an R/bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* 2016;**32**:289–91. https://doi.org/10.1093/bioinformatics/btv562
42. Hammal F, de Langen P, Bergon A *et al.* ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* 2022;**50**:D316–25. https://doi.org/10.1093/nar/gkab996
43. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**:R83. https://doi.org/10.1186/gb-2012-13-10-r83
44. Karimzadeh M, Ernst C, Kundaje A *et al.* Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* 2018;**46**:e120.
45. Yamada Y, Watanabe M, Yamanaka M *et al.* Aberrant methylation of the vascular endothelial growth factor receptor-1 gene in prostate cancer. *Cancer Sci* 2003;**94**:536–9. https://doi.org/10.1111/j.1349-7006.2003.tb01479.x
46. Baubec T, Schübeler D. Genomic patterns and context specific interpretation of DNA methylation. *Curr Opin Genet Dev* 2014;**25**:85–92.
47. Yin Y, Morgunova E, Jolma A *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 2017;**356**:eaaj2239. https://doi.org/10.1126/science.aaj2239
48. Wang T, Song J, Qu M *et al.* Integrative epigenome map of the normal Human prostate provides insights into prostate cancer predisposition. *Front Cell Dev Biol* 2021;**9**:723676.
49. Agarwal SK, Lee Burns A, Sukhodolets KE *et al.* Molecular pathology of the MEN1 gene. *Ann NY Acad Sci* 2004;**1014**:189–98. https://doi.org/10.1196/annals.1294.020
50. Van Der Heul-Nieuwenhuijsen L, Dits NF, Jenster G. Gene expression of forkhead transcription factors in the normal and diseased human prostate. *BJU Int* 2009;**103**:1574–80. https://doi.org/10.1111/j.1464-410X.2009.08351.x
51. Katoh M, Igarashi M, Fukuda H *et al.* Cancer genetics and genomics of human FOX family genes. *Cancer Lett* 2013;**328**:198–206. https://doi.org/10.1016/j.canlet.2012.09.017
52. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;**21**:5400–13. https://doi.org/10.1038/sj.onc.1205651
53. Issa J-P. Issa JP.CpG island methylator phenotype in cancer. *Nat Rev Cancer* 2004;**4**:988–93. https://doi.org/10.1038/nrc1507
54. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. *Trends Genet* 2021;**37**:1012–27. https://doi.org/10.1016/j.tig.2021.05.002
55. Heery R, Schaefer MH. DNA methylation variation along the cancer epigenome and the identification of novel epigenetic driver events. *Nucleic Acids Res* 2021;**49**:12692–705. https://doi.org/10.1093/nar/gkab1167
56. Mcheik S, Aptecar L, Coopman P *et al.* Dual role of the PTPN13 tyrosine phosphatase in cancer. *Biomolecules* 2020;**10**:1659. https://doi.org/10.3390/biom10121659
57. Meiers I, Shanks JH, Bostwick DG. Glutathione S-transferase pi (GSTP1) hypermethylation in prostate cancer: review 2007. *Pathology (Phila)* 2007;**39**:299–304. https://doi.org/10.1080/00313020701329906
58. Zhou W, Dinh HQ, Ramjan Z *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* 2018;**50**:591–602. https://doi.org/10.1038/s41588-018-0073-4
59. Hansen KD, Timp W, Bravo HC *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011;**43**:768–75. https://doi.org/10.1038/ng.865
60. Bock C, Beerman I, Lien W-H *et al.* DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol Cell* 2012;**47**:633–47. https://doi.org/10.1016/j.molcel.2012.06.019
61. Schultz MD, He Y, Whitaker JW *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 2015;**523**:212–6. https://doi.org/10.1038/nature14465
62. Angermueller C, Clark SJ, Lee HJ *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 2016;**13**:229–32. https://doi.org/10.1038/nmeth.3728

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

63. Bernstein BE, Mikkelsen TS, Xie X *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006;**125**:315–26. https://doi.org/10.1016/j.cell.2006.02.041
64. Azuara V, Perry P, Sauer S *et al.* Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* 2006;**8**:532–8. https://doi.org/10.1038/ncb1403
65. Kumar D, Cinghu S, Oldfield AJ *et al.* Decoding the function of bivalent chromatin in development and cancer. *Genome Res* 2021;**31**:2170–84. https://doi.org/10.1101/gr.275736.121
66. Kundaje A, Meuleman W, Ernst J *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30. https://doi.org/10.1038/nature14248
67. Meissner A, Mikkelsen TS, Gu H *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**:766–70. https://doi.org/10.1038/nature07107
68. Ohm JE, McGarvey KM, Yu X *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* 2007;**39**:237–42. https://doi.org/10.1038/ng1972
69. Rodriguez J, Muñoz M, Vives L *et al.* Bivalent domains enforce transcriptional memory of DNA methylated genes in cancer cells. *Proc Natl Acad Sci USA* 2008;**105**:19809–14. https://doi.org/10.1073/pnas.0810133105
70. Hinoue T, Weisenberger DJ, Lange CPE *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012;**22**:271–82. https://doi.org/10.1101/gr.117523.110
71. Bernhart SH, Kretzmer H, Holdt LM *et al.* Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep* 2016;**6**:37393. https://doi.org/10.1038/srep37393
72. Kretzmer H, Bernhart SH, Wang W *et al.* DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat Genet* 2015;**47**:1316–25. https://doi.org/10.1038/ng.3413
73. Xie F, Armand EJ, Yao Z *et al.* Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. *Cell Genomics* 2023;**3**:100342. https://doi.org/10.1016/j.xgen.2023.100342

Downloaded from https://academic.oup.com/nar/article/53/18/gkaf949/8268404 by European Institute of Oncology user on 27 May 2026

Received: February 1, 2024. **Revised:** July 12, 2025. **Accepted:** August 23, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.