




A method of sample-wise region-set enrichment analysis for DNA methylomics

Ryu Minegishi^{‡,1}, Osamu Gotoh^{‡,1} , Norio Tanaka¹, Reo Maruyama², Jeffrey T Chang³ & Seiichi Mori^{*,1}

¹Project for Development of Innovative Research on Cancer Therapeutics, Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan

²Project for Cancer Epigenomics, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan

³Department of Integrative Biology & Pharmacology, University of Texas Health Science Center, Houston, TX 77030, USA

*Author for correspondence: seiichi.mori@jfcrc.or.jp

‡Authors contributed equally

Aim: Gene set analysis has commonly been used to interpret DNA methylome data. However, summarizing the DNA methylation level of a gene is challenging due to variability in the number, density and methylation levels of CpG sites, and the numerous intergenic CpGs. Instead, we propose to use region sets to annotate the DNA methylome. **Methods:** We developed single sample region-set enrichment analysis for DNA methylome (methyl-ssRSEA) to conduct sample-wise, region-set enrichment analysis. **Results:** Methyl-ssRSEA can handle both microarray- and sequencing-based platforms and reproducibly recover the known biology from the methylation profiles of peripheral blood cells and breast cancers. The performance was superior to existing tools for region-set analysis in discriminating blood cell types. **Conclusion:** Methyl-ssRSEA offers a novel way to functionally interpret the DNA methylome in the cell.

Lay abstract: Gene set analysis has been a common way to understand the meaning of DNA methylome data. However, organizing the DNA methylation level of a gene is challenging due to variation in the number, density and extent of methylation, of methylation sites, and the substantial number of methylation sites between genes. Instead, we propose to use region sets for the organization. We developed single sample region-set enrichment analysis for DNA methylome (methyl-ssRSEA) to conduct region-set analysis for every sample. Methyl-ssRSEA can handle both microarray- and sequencing-based methods and repeatedly find the known characters from the methylation patterns of peripheral blood cells and breast cancers. The performance was better than existing tools for region-set analysis in differentiating blood cell types. Methyl-ssRSEA offers a novel way to find the features of DNA methylome in the cell.

First draft submitted: 18 February 2021; Accepted for publication: 21 June 2021; Published online: 7 July 2021

Keywords: DNA methylome • functional annotation • gene set analysis • region set analysis • sample wise analysis • transcription factor binding site

In multicellular organisms, DNA methylation is a critical epigenetic modification that drives and maintains various processes, such as cell differentiation, sex differentiation, aging and the development of disease. A well-studied consequence of DNA methylation is the modulation of transcriptional activity [1,2]. DNA methylation at the promoter and more distal regulatory elements plays an important role in gene silencing via multiple distinct mechanisms, including the inhibition of transcription factor (TF) binding, the modification of chromatin opening and the recruitment of methyl CpG binding proteins [1,2]. Previous studies have demonstrated that cell type-specific hypomethylated genomic regions are enriched in TF-binding sites (TF-BSs) known to be active in that cell type [2,3]. Therefore, DNA methylomic profiles can provide clues to the TFs and regulatory elements that are active in a cell type of interest.

Whereas a wealth of statistical methodologies is available to analyze transcriptomic data, such as RNA-seq and expression microarray, there are only a few options available to interpret the biology of the DNA methylation data [1,4–12]. Many of the existing approaches are relatively straightforward translations of ones developed for RNA-

Seq. For instance, one commonly adapted method is Gene Set Enrichment Analysis (GSEA), which determines whether the expression of a pre-defined gene set representing a biological pathway or process, such as cell proliferation or Ras pathway activation, is significantly different between two groups of samples representing distinct conditions or phenotypes [13]. A limitation of GSEA is that it only scores the differences in activity between two groups and does not score pathway activity in single samples. However, this was addressed in the follow-up algorithm, single-sample GSEA (ssGSEA) [14].

Versions of gene-set analyses have been used to analyze DNA methylome data [7–12,15], but the adaptation often ignores distinct features including the differential number and density of CpG sites within a gene [7–10], the differential effect of CpG methylation per gene structure [11,12] and the existence of a substantial number of intergenic CpGs [1,2]. In seeking to overcome this limitation, several methods have been developed to analyze genomic regions (rather than genes); however, aside from one, none of these tools score region activity in single samples [16–18]. Here, we describe the development of single sample region-set enrichment analysis for DNA methylome (Methyl-ssRSEA), a methodology in which functionally annotated genomic regions are used to assess the DNA methylation status of CpG sites, circumventing the need to summarize the methylation status of genes. Importantly, Methyl-ssRSEA provides a score per sample that can be utilized for further statistical analyses. We describe and evaluate our methodology using peripheral blood and breast tumor samples – biological models whose cell states are relatively well established.

Methods

Development of a tool for sample-wise region-set enrichment analysis of the DNA methylome

The previously established framework of GSEA requires a summary of the β values across the CpG sites for a gene. However, this is challenging due to the variable number and density of CpG sites, the variable degree of their methylation within a gene [7–10] and the difficulty in assigning intergenic CpGs to a specific gene (Figure 1A [1,2]). We therefore sought to link genomic regions directly to cell states (Figure 1B) and developed a tool for sample-wise region-set analysis, hereafter Methyl-ssRSEA (Figure 1C). A region set is a list of genomic regions linked to an annotation, such as TF binding or histone mark. Methyl-ssRSEA compares the region set against the methylation profile of a sample of interest and produces a score that quantifies the overall methylation of the region set in that sample (Figure 1C).

Algorithm used in Methyl-ssRSEA

The algorithm used in Methyl-ssRSEA is derived from the one for ssGSEA [14]: whereas ssGSEA works with expression values from a gene set, Methyl-ssRSEA uses the β values for CpG sites from a region set. ssGSEA – either the initial version described in the original manuscript [14], or the most updated version implemented as a webtool (www.genepattern.org/modules/docs/ssGSEAProjection/4) – utilizes the rank-normalized expression values not the z-scores as the input data type. Accordingly rank-normalized β values for CpG sites from a region set are used for Methyl-ssRSEA. Differentially methylated regions (DMRs) are not selected during the process of Methyl-ssRSEA. All β values within TF-BSs are simply used for enrichment analysis. In ssGSEA, an enrichment score is computed for a gene set using an Empirical Cumulative Distribution Function (ECDF) of rank-normalized expression values in the gene set compared with that in the background set of genes. Instead, Methyl-ssRSEA calculates an enrichment score for a region set by comparing an ECDF of rank-normalized β values of CpG sites included in the region set against those for all remaining CpG probes.

The entire set of CpG sites on a DNA methylation array (or in whole-genome bisulfite sequencing [WGBS] data, or in reduced representation bisulfite sequencing [RRBS] data) is represented by $U = \{p_1, p_2, \dots, p_N\}$, where N is total number of CpGs. U is sorted such that p_1 is the CpG site with the largest β value, and p_N has the smallest. The rank of a CpG site p^j is represented by $|p_j|$. The CpG sites in a region set is defined R , where $R \subset U$. An enrichment score, $ES(R)$, is subsequently computed from the cumulative sum of differences between a weighted ECDF $P_{in}(R, i)$ for CpGs included in a region set and an ECDF $P_{out}(R, i)$ for CpGs not included in the region set. That is,

$$ES(R) = \sum_{i=1}^N [P_{in}(R, i) - P_{out}(R, i)]$$

Here, $P_{in}(R, i)$ and $P_{out}(R, i)$ are formularized as below.

$$P_{in}(R, i) = \frac{\sum_{p_j \in R, |p_j| \geq |p_i|} |p_j|^\alpha}{\sum_{p_j \in R} |p_j|^\alpha}$$

$$P_{out}(R, i) = \sum_{p_j \in U \setminus R, |p_j| \geq |p_i|} \frac{1}{(N - N_R)}$$

The number of CpGs included in the region set is indicated by N_R . An exponent α in $P_{in}(R, i)$ can be tuned to calibrate the weight of the rank of the regions in a region set, as $\alpha = 0$ (no weight), 0.25 (partial weight) or 1 (full weight) in the ssGSEA algorithm.

Use of methyl-ssRSEA 1: components

Methyl-ssRSEA requires three inputs: one or more region sets, the test data and CpG coordinates. A region set is represented as a list of genomic coordinates (chromosome name, start and end position) in Browser Extensible Data format. The test data is a matrix of β values of CpG sites for each sample. The M value can be also used in the test data. The resultant enrichment score will be same regardless of the type of value. Finally, the CpG coordinates describe the genomic coordinate (chromosome name and genomic position) for each CpG site (Figure 1C). To reflect the weight for the rank of a gene set, we used $\alpha = 1$ in the current study.

Use of methyl-ssRSEA 2: region-set data

In the current study, we defined regions from TF-BSs or histone marks. We obtained TF-BSs from ENCODE (ver. 94) (www.encodeproject.org/) [21], ReMap (<http://remap.univ-amu.fr/>) [22] and CODEX (<http://codex.stemcells.cam.ac.uk/>) [23]. Whereas ENCODE provides genomic region data generated by the consortium, ReMap and CODEX publish the curated and processed data from Gene Expression Omnibus and ArrayExpress, which are not included in ENCODE [21–23].

Use of methyl-ssRSEA 3: DNA methylome data

We used DNA methylomes of peripheral blood cell types (GSE103541) [24] and The Cancer Genome Atlas (TCGA) breast cancers [25] (Illumina MethylationEPIC and Methylation450 arrays). We downloaded the data from the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) and from the Genomic Data Commons Data Portal (<https://portal.gdc.cancer.gov/>). The downloaded IDAT files were converted to β -value matrices after background correction and control normalization with minfi in R [26]. For WGBS and RRBS data obtained from ENCODE, we used the ‘methylation state at CpG’ data from the bedMethyl file generated by Bismark, which is equivalent to the β values from Illumina microarrays [21]. Non-numeric or empty β values in the test data matrix were ignored. We used human genome assembly hg19 for the Illumina microarray-only analyses (for Figures 2 & 3); and hg38 for the comparative analyses between Illumina microarray, WGBS and RRBS data (Figure 5). For the RRBS data, originally prepared with the hg19 genome in the ENCODE project [21], the reference was converted from hg19 to hg38 by liftOver in R (ver. 1.14.0); (Figure 5). Methyl-ssRSEA was implemented in R, and a detailed protocol can be found in the Supplementary Document (Methyl-ssRSEA.R: User Guide).

Proof-of-concept experiment 1: utility of Methyl-ssRSEA in interpreting blood-cell DNA methylomes

We first sought to test whether Methyl-ssRSEA could recover the methylation patterns seen in cell differentiation by running the test on the DNA methylome of peripheral blood cells (Figure 2). The test data included DNA methylation profiles for five peripheral blood cell types purified from 28 individuals (total 140 samples): B cells, CD4⁺ T-cells, CD8⁺ T-cells, granulocytes and monocytes [24]. For the region-set data, we compiled 1884 region-sets derived from ChIP-seq data from the ENCODE (ver. 94) [21] and CODEX [23] databases, including 1840 ChIP-seq profiles derived from multiple blood cell types, such as leukemia or lymphoma cell lines, immortalized lymphoblasts and normal primary hematopoietic cells.

Figure 2A shows a representative example of how Methyl-ssRSEA scored the DNA methylome in a B-cell sample from one individual. The β values at 866,836 CpG sites on the Illumina array were summarized to one ssRSEA score based on the genomic regions bound by BCL6 (a transcription factor relevant for B-cell differentiation) in

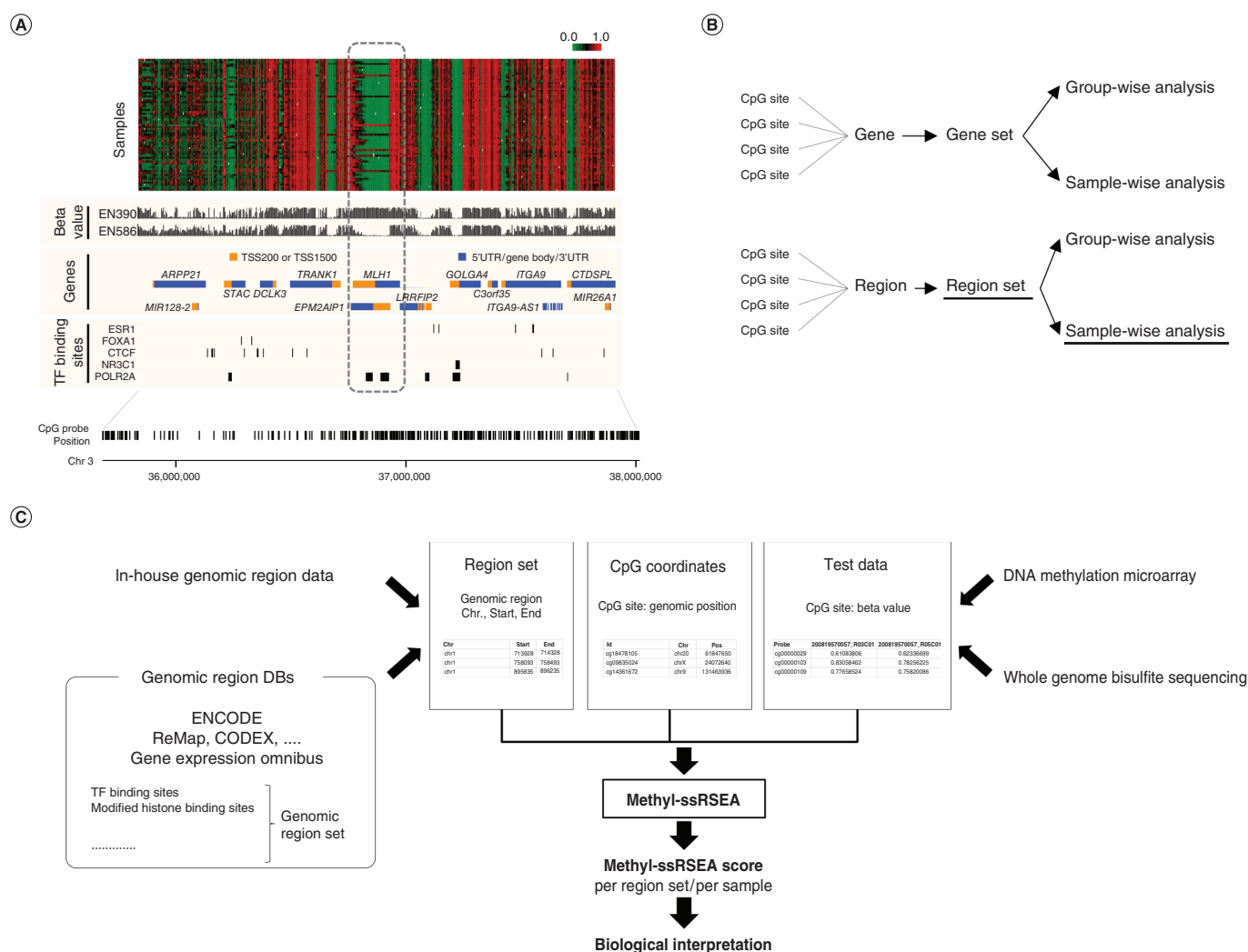


Figure 1. Sample-wise region-set enrichment analysis of the DNA methylome. (A) An example of the relationship between methylation status of CpG sites, and genes or TF-Bs in the 2.3-Mb genomic region surrounding the *MLH1* gene (35,680,290–38,017,188 on chromosome 3 in GRCh37). Panels from top to bottom: Heatmap of β values in 69 endometrial cancer samples [19] detected by Illumina Infinium Methylation EPIC array; bar plots of the β values in EN390 and EN586 endometrial cancers with *MLH1* promoter hyper- and hypo-methylation; positions of genes with annotations of promoter (TSS200 or TSS1500) and 5' UTR/gene body/3' UTR; binding sites of TFs in Ishikawa endometrial cancer cell line (ENCODE ver. 94); and CpG probe positions on the Infinium Methylation EPIC array. The rounded dashed rectangle indicates the surrounding region of the *MLH1* gene. Annotations for gene structure were provided by Illumina based on the canonical transcript [20]. Note that annotations of genes and TF-Bs are shown only for CpG sites included in the array. **(B)** Methodological concepts in summarizing DNA methylome data. Underlines indicate the measures adopted in the current study. **(C)** A scheme to perform Methyl-ssRSEA. A 'region set' is defined as a set of functionally annotated genomic regions, such as a set of binding sites of a TF or a histone mark, in the genome of a cell line, primary cells, or a tissue. In-house genomic region-set data or data from genomic region-set DB (e.g., ENCODE [21], ReMap [22] and CODEX [23]) can be used for the analysis. A region set comprises genomic coordinate information. Test data have β values for CpG sites per sample and can be derived from DNA methylation microarray and/or WGBS data. The CpG coordinate file contains CpG probe ID (or CpG site ID made from genomic position in case of WGBS data) with positional information. Methyl-ssRSEA is run on region set data and test data, with the CpG coordinate file as a bridge, and outputs of ssRSEA score per region set per sample.

DB: Database; Methyl-ssRSEA: Single sample region-set enrichment analysis for DNA methylome; ssRSEA: single sample region-set enrichment analysis; TF: Transcription factor; TF-Bs: Transcription factor binding sites; TSS: Transcriptional start site; WGBS: Whole-genome bisulfite sequencing.

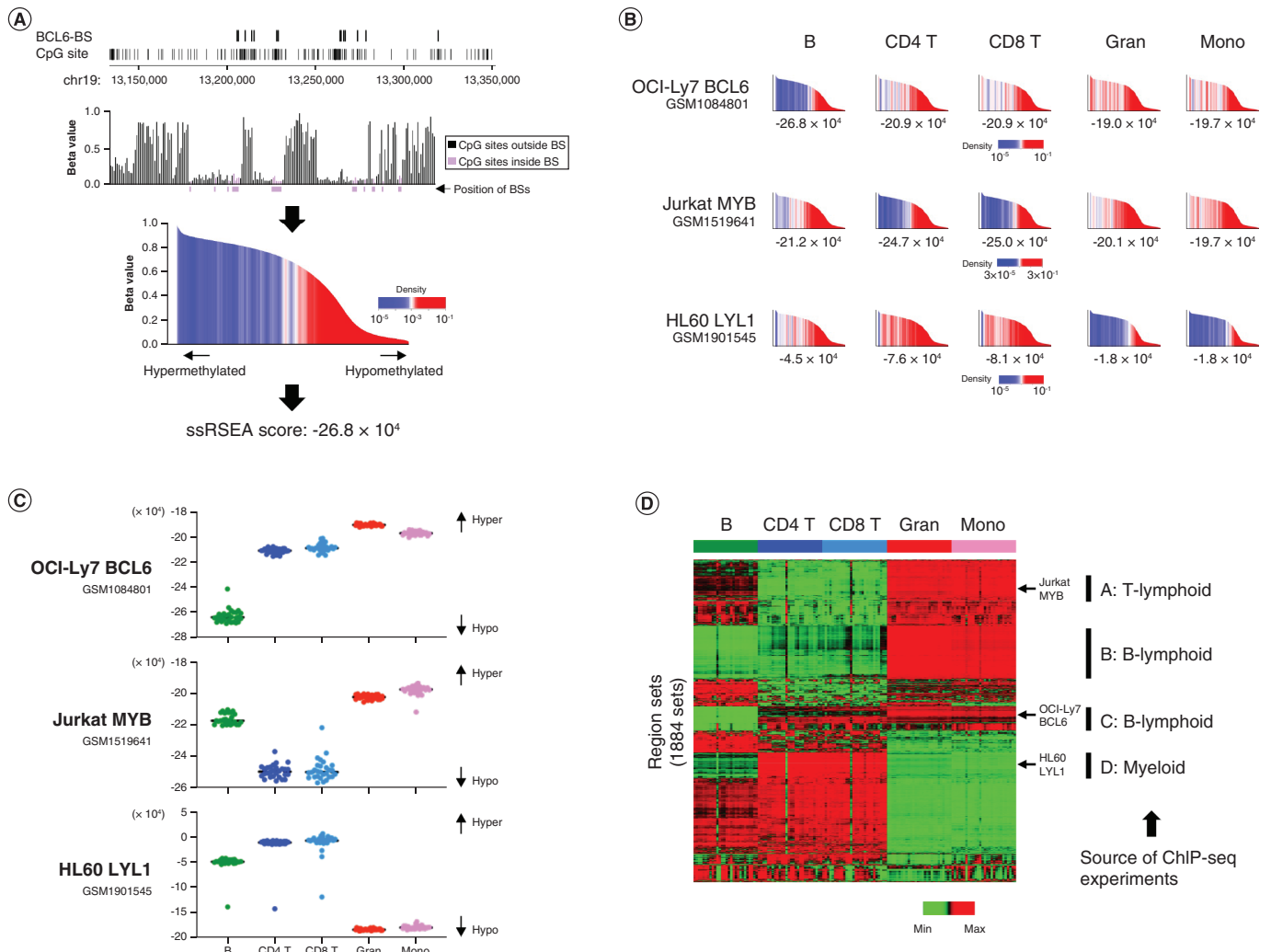


Figure 2. DNA methylation status of various TF-BSs in blood cells. (A) The process to compute a Methyl-ssRSEA score. Upper panel: Positions of BCL6-BSs and Illumina microarray CpG sites in the 216.3-kb genomic region (13,133,365–13,349,725 on chromosome 19). The region contains 11 BCL6-BSs and 198 CpG probes. Middle panel: Bar plots of β values in a peripheral B-cell sample from Person #1. Black and light purple bars represent β values for CpG probes outside and inside the BCL6-BSs, respectively. Positions of BCL6-BSs are also shown below the bar plots in light purple. Lower panel: Density plots of BCL6-BSs in the methylome of a B-cell sample. CpG sites are sorted according to the β values as hyper methylated (β -value = 1) to hypomethylated (β -value = 0). Density of CpGs included in BCL6-BSs over surrounding 2000 CpG sites (1000 hyper- and 1000 hypo-methylated) was calculated, and the moving averages are shown as a heatmap. Blue and red colors indicate sparse and dense distribution of CpG sites within BCL6-BSs, respectively. Note that BCL6-BSs are rarely seen in hypermethylated DNA regions in B cells. (B) Density plots of BCL6-, MYB- and LYL1-BSs in five peripheral blood cell types of Person #1. The three TF-BSs were derived from ChIP-seq data using OCI-Ly7, Jurkat, and HL60 cell lines. A computed Methyl-ssRSEA score is shown below the density heatmap. (C) ssRSEA scores in five blood cell types. ssRSEA scores for BCL6-, MYB- and LYL1-BSs were computed for B cells, CD4 T cells, CD8 T cells, granulocytes and monocytes collected from 28 persons. (D) Heatmap for ssRSEA scores of 1884 region sets in five peripheral blood cell types. ssRSEA scores are presented as a heatmap after centering and normalization for samples and region sets. Samples were aligned according to cell type; region sets were clustered with hierarchical clustering. The positions for OCI-Ly7 BCL6-, Jurkat MYB-, and HL60 LYL1-BSs are indicated by arrows on the right. Also on the right are the region-set clusters (A–D) of TF-BSs in T-lymphoid, B-lymphoid, and myeloid cell lines or primary cells. Red and green colors indicate hyper- and hypo-methylation of a region set in a sample. BS: Binding site; ChIP-seq: Chromatin Immunoprecipitation Sequencing; Gran: Granulocyte; Mono: Monocyte; ssRSEA: single sample region-set enrichment analysis; TF: Transcription factor; TF-BSs: Transcription factor binding sites.

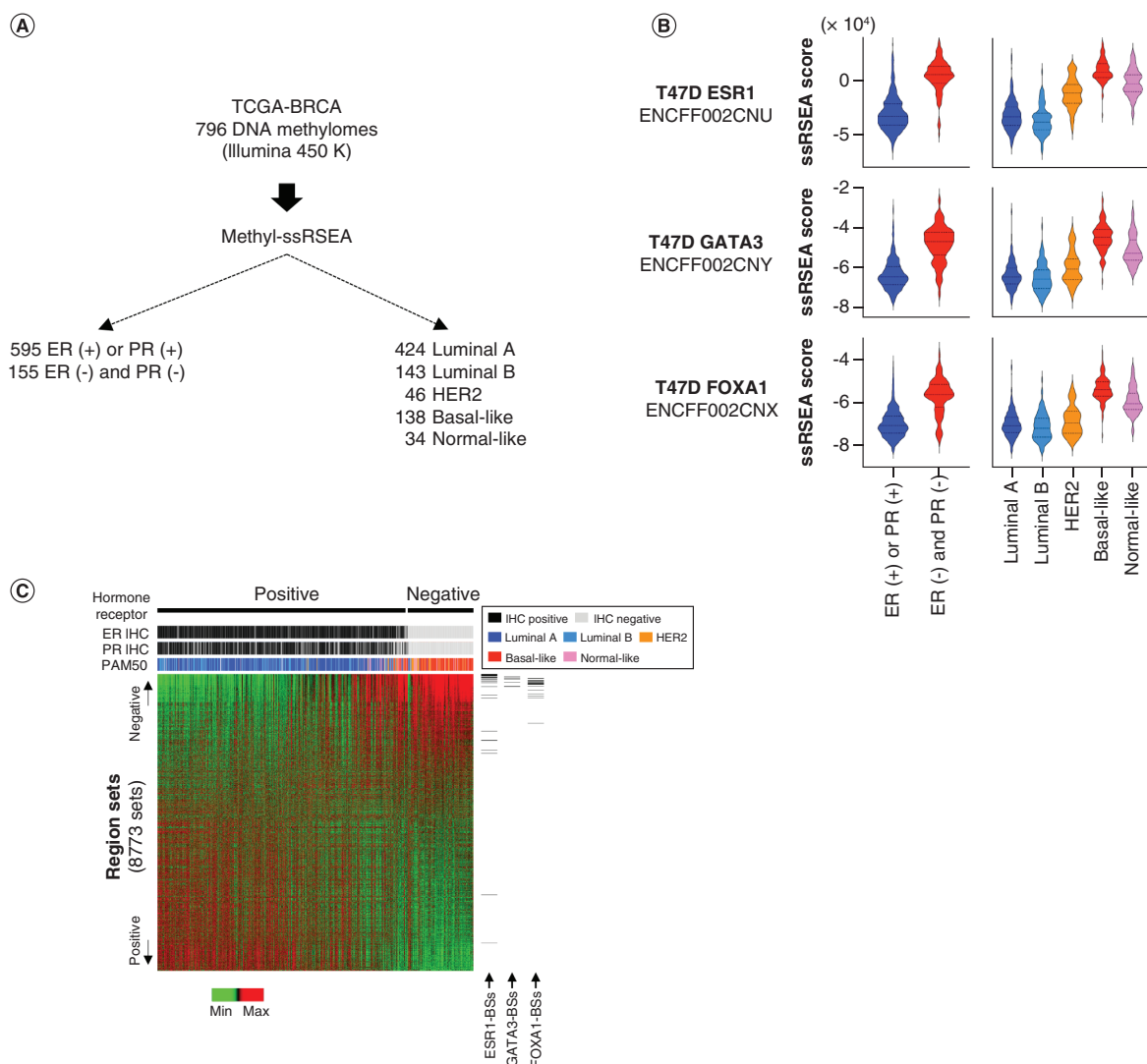


Figure 3. DNA methylation status of ER related TF-BSs in TCGA breast cancer samples. (A) Summary of TCGA breast cancer DNA methylome data. TCGA DNA methylome was measured with the Illumina 450 K array. We show the number of samples with information for hormonal status or molecular subtype. **(B)** ssRSEA scores of ESR1-, GATA3- and FOXA1-BSs for hormonal status or molecular subtype. T47D is a breast cancer cell line positive for ER and PR. **(C)** Hormonal status-associated DNA methylation of TCGA breast cancer. Heatmap of ssRSEA scores for 8,773 region sets is shown. Pairwise comparisons were performed with SAM on ssRSEA scores regarding hormone receptor positivity. Region sets were sorted according to *D*-values of SAM. To sort samples, we used ssRSEA scores of the top region set (ESR1-BSs: E2 + P4 treated breast cancer xenograft replicate 1 [GSE93108 from ReMap]), which was tightly correlated with hormonal receptor positivity. Red and green colors indicate hyper- and hypo-methylation of a region set in a sample. Positivity for ER/PR IHC and PAM50 molecular subtypes are shown with color codes above the heatmap. The positions of the region sets for 293 ESR1-, 47 GATA3-, and 152 FOXA1-BSs are shown using ticks on the right. BRCA: Breast cancer; BS: Binding site; (+): positive; (-): negative; ER: Estrogen receptor; IHC: Immunohistochemistry; PR: Progesterone receptor; SAM: Significance analysis of microarray; ssRSEA: single sample region-set enrichment analysis; TCGA: The Cancer Genome Atlas; TF-BSs: Transcription factor binding sites.

OCI-Ly7 cells (a diffuse large B-cell lymphoma cell line) (GSM1084801 from CODEX). The 216.3-kb genomic region (chromosome 19, bases 13,133,365–13,349,725) contained 11 OCI-Ly7 BCL6-BSs (binding sites) and 198 Illumina CpG sites. Within these BSs, 22 CpG probes were hypomethylated, whereas a variable range of methylation was found on the 176 CpG probes outside the BCL6-BSs. Expansion of the analysis to the entire array showed that OCI-Ly7 BCL6-BSs are consistently hypomethylated (Figure 2A). After rank normalization of the β

values, the enrichment score was computed using the empirical cumulative distribution functions of the CpG sites inside and outside the region set (Figure 2A).

We subsequently performed Methyl-ssRSEA for the four remaining cell types (CD4⁺ T, CD8⁺ T, granulocytes and monocytes) and for two additional region sets (MYB-BSs in Jurkat, an acute T-cell leukemia cell line; and LYL1-BSs in HL60, an acute myelogenous leukemia cell line); (Figure 2B). The analyses revealed that OCI-Ly7 BCL6-BSs, Jurkat MYB-BSs and HL60 LYL1-BSs exhibited hypomethylation in B cells, T cells (CD4⁺ and CD8⁺ T-cells) and myeloid cells (granulocytes and monocytes), respectively, in the density plots as well as in the enrichment scores (Figure 2B). This recapitulates the lineages of these cell lines: OCI-Ly7, Jurkat and HL60 cells are B-lymphoid, T-lymphoid, and myeloid cells, respectively. This finding is consistent with the concept that active TFs bind to open hypomethylated genomic regions in a cell-type specific manner [2,3]. This relation was reproducibly observed in the blood-cell DNA methylomes of 27 additional individuals (Figure 2C). Clustering Methyl-ssRSEA scores for 1884 region sets revealed multiple clusters, including four region-set clusters: A–D (Figure 2D & Supplementary Table 1). Region-set cluster A exhibited hypomethylation in CD4⁺/CD8⁺ T cells and contained the TF-BSs and active histone marks for T-lymphoid cells, which include Jurkat, MOLT-3, CUTLL1 and thymus (MOLT-3 and CUTLL1 cell lines are derived from acute T-lymphoblastic leukemia). The TF-BSs from B-lymphoid cells, such as GM12878 (an EBV-transformed lymphoblast cell line), MM1.S (a multiple myeloma cell line), and Ramos (a Burkitt lymphoma cell line) were hypomethylated and are included in cluster B. Region-set cluster C, with hypomethylation in B cells, included the TFs from B-lymphoid OCI-Ly7 and Raji (a Burkitt lymphoma cell line). The TF-BSs and active histone marks from CD14⁺ monocytes, peripheral blood-derived macrophages, HL60, Kasumi-1 (an acute myelogenous leukemia cell line), and other myeloid cells were grouped as cluster D and found to be hypomethylated (Figure 2D). Collectively, these results demonstrate that Methyl-ssRSEA can uncover a clear relationship between DNA hypomethylated regions and TF-BSs in blood cell types.

Proof-of-concept experiment 2: DNA methylome & hormonal status in breast cancer

A previous study identified DNA hypo- and hyper-methylation of ESR1-, GATA3- and FOXA1-BSs in estrogen receptor (ER)-positive and ER-negative breast cancer lineages [27]. As a second proof-of-concept analysis, we sought to relate the DNA methylation pattern revealed by Methyl-ssRSEA with the hormonal status of breast cancer (Figure 3). As the test data, we used the TCGA breast cancer DNA methylome data (ver. 10.1) [25], which were profiled with Illumina HumanMethylation450 microarrays. In total, 796 DNA methylome samples were available, including profiles from 595 ER- or progesterone receptor (PR)-positive tumors, and 155 ER- and PR-negative tumors. In terms of PAM50 molecular subtypes, there were 424 luminal A, 143 luminal B, 46 HER2, 138 basal-like, and 34 normal-like breast cancers (Figure 3A). For Methyl-ssRSEA, we used a total of 8773 genomic region sets obtained from ENCODE (ver. 94) and ReMap databases [22], which included 293 ESR1-, 47 GATA3-, and 133 FOXA1-BSs from various ChIP-seq experiments across 359 breast and 114 non-breast cancer cell lines or tissues. Figure 3B shows the Methyl-ssRSEA scores for ESR1-, GATA3-, and FOXA1-BSs of T47D, an ER-positive breast cancer cell line. Hormone receptor-positive breast cancer samples exhibited hypomethylation of the ESR1-, GATA3- and FOXA1-BSs, which is consistent with previous observations [27]. Likewise, luminal A or luminal B breast cancers – 98.5% of which were positive for ER or PR showed – hypomethylation in these three TF-BSs in a PAM50-based subtyping scheme (Figure 3B).

Next, we performed a comprehensive evaluation of all 8773 region sets. The region-set enrichment scores of the TCGA DNA methylomes of hormone receptor-positive and -negative samples were compared using Significance Analysis of Microarray (SAM) with R [28]. The results shown in Figure 3C confirm that ESR1-, GATA3- and FOXA1-BSs are indeed hypomethylated in hormone receptor-positive breast cancer samples, with $p < 0.0001$ for all three BSs by hypergeometric tests (Figure 3C & Supplementary Table 2). These observations indicate high reproducibility of the Methyl-ssRSEA results in relation to hormonal receptor positivity in TCGA breast cancer DNA methylome.

Methyl-ssRSEA & other available tools: feature comparison

We searched PUBMED and Google Scholar on 11 January 2021, and found three computational tools that could perform region set analysis on DNA methylomes: Locus Overlap Analysis (LOLA) [17], Methylation-based Inference of Regulatory Activity (MIRA) [16] and COordinate COvariation Analysis (COCO) of epigenetic heterogeneity [18]. In Table 1, we provide a detailed comparison of the features of these tools against Methyl-ssRSEA. Aside from differences in the statistical approach, the most significant difference is the strategy employed

Table 1. Computational tools to evaluate DNA methylome signals per region set.										
Tool	Evaluation per region set	Statistical basis	Use of off-region signals		Computation		Requirement of DMRs	Programming language	Study (year)	Ref.
			Genome-wide	Local surrounding	Sample-wise computation	Group-wise computation				
Methyl-ssRSEA	Yes	ECDF	Yes	No	Yes	No	No	R	Current study	
LOLA	Yes	Fisher	Yes	No	No	Yes	Yes	R	Sheffield (2016)	[17]
MIRA	Yes	Average	No	Yes	Yes	No	No	R	Lawson (2018)	[16]
COCOA	Yes	Matrix factorization	No	No	No	No	No	R	Lawson (2020)	[18]
COCOA: Coordinate Covariation Analysis of epigenetic heterogeneity; DMR: Differentially Methylated Region; ECDF: Empirical Cumulative Distribution Function; LOLA: Locus Overlap Analysis; MIRA: Methylation-based Inference of Regulatory Activity; Methyl-ssRSEA: Single sample Region Set Enrichment Analysis for DNA Methylome.										

to generate a background distribution. Methyl-ssRSEA and LOLA use information across the genome, whereas MIRA uses genomic locations surrounding the region set. COCOA does not use any information outside the region set. As to the output, Methyl-ssRSEA and MIRA provide a score for each sample, and LOLA was developed for group-wise comparison. COCOA does not directly produce any score for a sample or for a group. Prior to the region-set analysis, LOLA requires the user to calculate differentially methylated regions (DMRs); (Table 1).

Methyl-ssRSEA & other available tools: performance comparison

We obtained the R code for LOLA [17], which was built in RnBeads (ver. 2.4.0) [6], and MIRA [16] from the Bioconductor repository (www.bioconductor.org/) [6,16,17]. Prior to LOLA analysis, we determined the DMRs using a cut-off automatically computed using RnBeads (ver. 2.4.0) [6].

We compared LOLA, MIRA and single sample region-set enrichment analysis (ssRSEA) with respect to their abilities to discriminate blood-cell type: B cells versus non-B cells (using the region set of OCI-Ly7 BCL6-BSs), T cells (CD4 and CD8 T cells) versus non-T cells (using the region set of Jurkat MYB-BSs), and myeloid cells (granulocytes and monocytes) versus non-myeloid cells (using the region set of HL60 LYL1-BSs); (Figures 2 & 4). Because of the similarities in the DNA methylome, CD4 and CD8 T cells, and granulocytes and monocytes were combined, respectively, as T cells and myeloid cells. We excluded COCOA from the comparison because it does not provide sample-specific or group scores. Fisher exact tests using R were used to evaluate the performance. Youden indices, computed after Receiver Operating Characteristic (ROC) analyses (using R) for MIRA and ssRSEA, were used as cut-off values for blood-cell discriminations. For the Fisher test for LOLA, the built-in function was used for computation. Fisher exact tests revealed that MIRA and Methyl-ssRSEA were best able to distinguish all three of the blood cell types (Figure 4A).

MIRA and Methyl-ssRSEA were further compared through sample-wise outputs with Spearman correlation (using GraphPad Prism; Figure 4B) and ROC analyses (using R; Figure 4C). Whereas scores derived from MIRA and Methyl-ssRSEA exhibited strong correlations for all three region sets (Figure 4B), the discriminating capability of Methyl-ssRSEA was slightly better than that of MIRA (Figure 4C). MIRA misclassified 1 and 2 blood-cell samples in B- and T-cell discriminations, but Methyl-ssRSEA classified all the samples correctly. The comparisons demonstrate that the scores from Methyl-ssRSEA are more accurate reflections of the state of the DNA methylome than other available tools.

DNA methylome from WGBS or RRBS

Although microarray has been widely used to profile the DNA methylome, sequencing-based technologies, such as WGBS and RRBS, have become increasingly popular and more available in public databases. To enhance the utility of Methyl-ssRSEA, we extended the framework to handle such sequencing-based data.

To compare ssRSEA scores between microarray and WGBS or RRBS data, we used the DNA methylome data from ENCODE ver. 94. WGBS and RRBS data files used in the comparison were derived from nine and six cell lines, respectively (Supplementary Table 3). Each cell line has a set of dual-color (green/red) microarray data files. The number of CpG data points used for this analysis were 58,304,912, 1,286,457 and 862,573 (on chromosomes 1–22, X and Y) for WGBS, RRBS and microarray, respectively. In the microarray, missing values were labeled as ‘NaN’ by minfi. In the WGBS and RRBS data, values at the CpG sites with sequencing depths below ten were considered missing [29]. Whereas the number of missing values in the MethylationEPIC array data was negligible

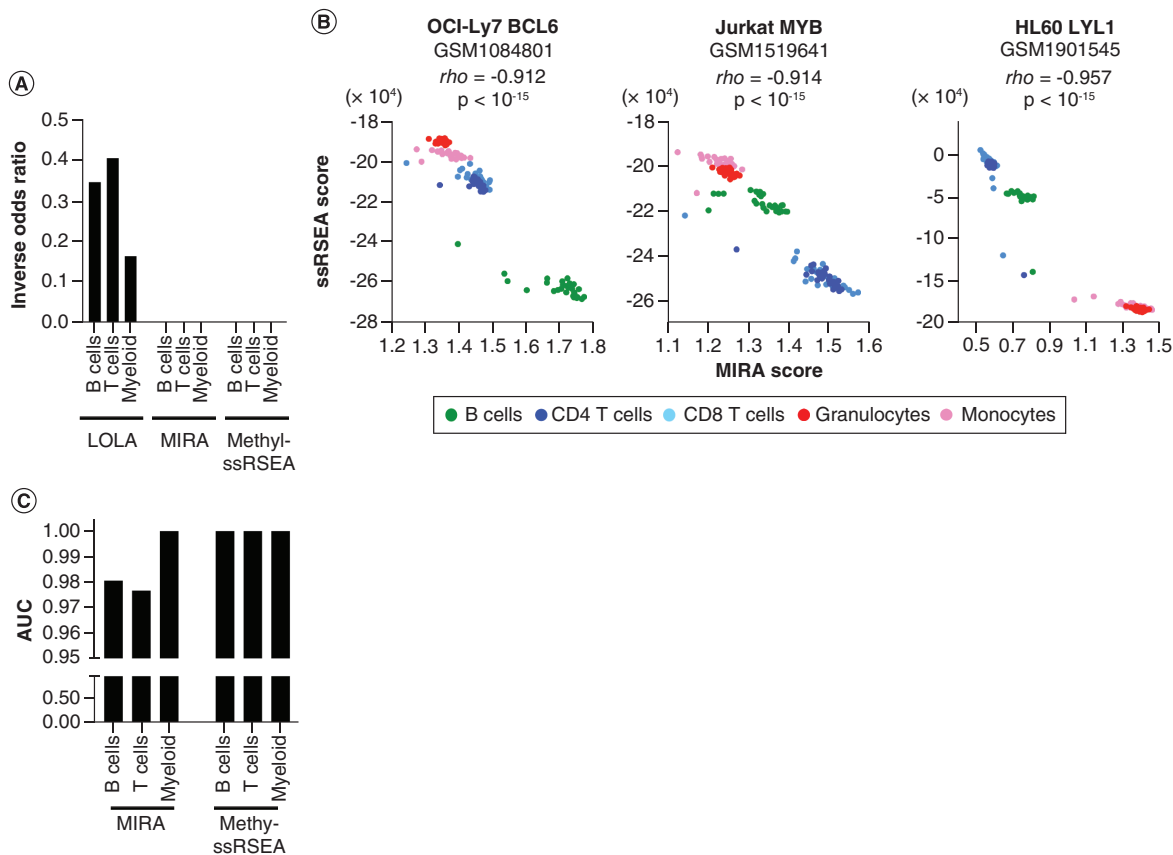


Figure 4. Comparison of computational tools to evaluate DNA methylome signals per region set. (A) Inverse odds ratio of Fisher exact tests in discriminating capability of blood cell types. Odds ratio for B cells versus non-B cells (using the region set of OCI-Ly7 BCL6-BSs), T cells versus non-T cells (using the region set of Jurkat MYB-BSs) and myeloid cells versus nonmyeloid cells (using the region set of HL60 LYL1-BSs) are shown in bar plots. Tools are indicated below blood-cell type names. Note that the inverse odds ratios for MIRA and Methyl-ssRSEA were zero. **(B)** Relationships between MIRA and Methyl-ssRSEA scores. MIRA and Methyl-ssRSEA scores for BCL6-, MYB-, and LYL1-BSs were computed for B cells, CD4 T cells, CD8 T cells, granulocytes and monocytes collected from 28 people and are presented as scatter plots. ρ and p -value were computed by Spearman correlation. **(C)** AUC in ROC analysis. Note that AUCs are 1 in myeloid discrimination by MIRA and in B cell, T cell, and myeloid discriminations by Methyl-ssRSEA. AUC: Area under the curve; BS: Binding site; LOLA: Locus Overlap Analysis; Methyl-ssRSEA: Single sample region set enrichment analysis for DNA methylome; MIRA: Methylation-based Inference of Regulatory Activity; ROC: Receiver operating characteristic.

(0.0016%–0.0085%; median 0.0034%), the number in WGBS and RRBS was considerably higher (26.65%–70.17%; median 37.89%, and 34.82%–85.40%; median 70.11%, respectively). For the comparison, the missing values from both platforms were filtered out and the remaining values were used in the subsequent analysis. The CpG coordinate file was generated from bedMethyl files from ENCODE (ver. 94) [21].

Methyl-ssRSEA was run using 4280 region sets of TF-BSs or modified histone-BSs from ENCODE. We show a representative 2D chart for IMR-90 scores from WGBS and microarray, which indeed exhibits tight correlation in the Spearman analysis (Figure 5A). The remaining cell lines also showed high Spearman coefficients between microarray and WGBS or RRBS data (Figure 5B & C), indicating consistency across different platforms and the ability to use Methyl-ssRSEA to analyze WGBS and RRBS data.

Discussion & conclusion

Gene set analysis can be used for the biological interpretation of DNA methylome data but requires gene-level summarization of multiple CpG sites [15]. Several methods have been developed to avoid the bias derived from the differential number and density of CpG sites in a gene: some methods adjust for the number of CpG sites associated with a gene by weighted resampling [7,8], by logistic kernel machine modeling [9] or by robust rank aggregation [10];

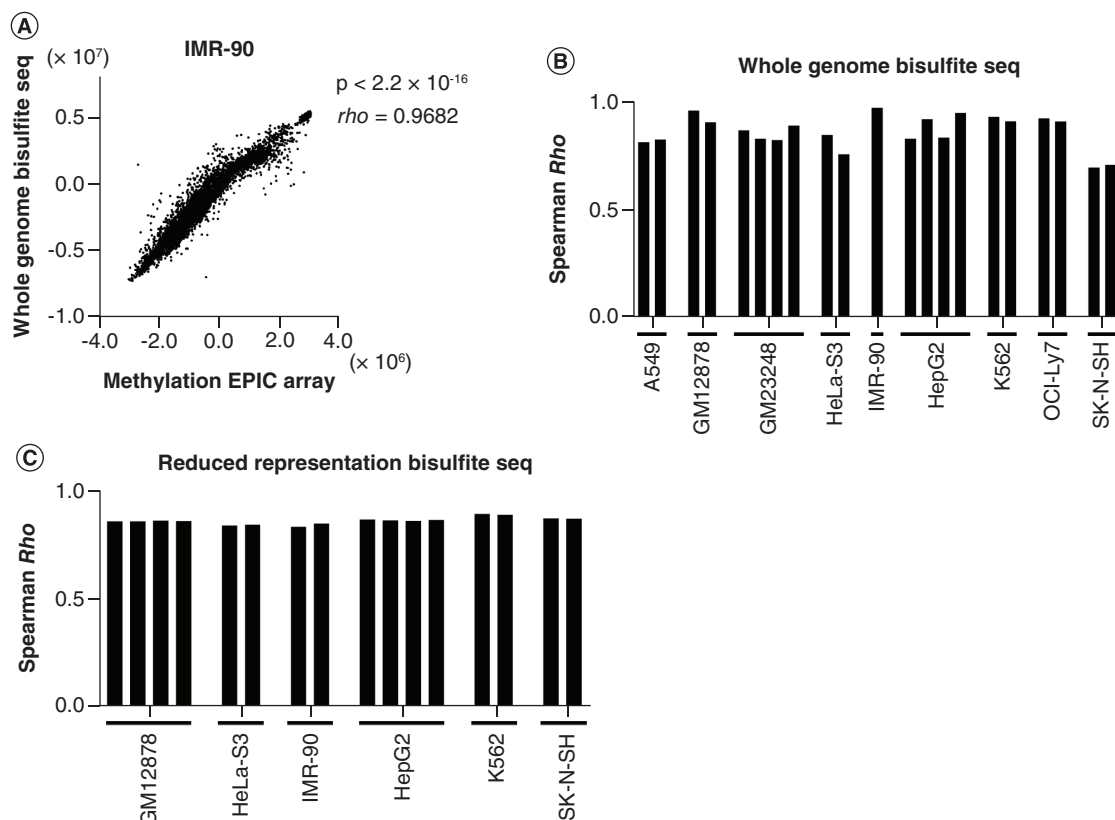


Figure 5. Comparisons of single sample region set enrichment analysis for DNA methylome results from DNA methylation microarray and whole-genome bisulfite sequencing or reduced representation bisulfite sequencing data. (A) Two-dimensional dot plots of enrichment scores in IMR-90 cell line. Enrichment scores were computed from Methylation EPIC array and WGBS data. The p -value and ρ were computed with Spearman correlation. **(B)** Spearman's rank correlation coefficients (ρ) of enrichment scores derived from DNA methylation microarray and WGBS data in nine cell lines. Cell line names are shown below the bar plots. Note that there are multiple correlation coefficients for biological or technical replicates of WGBS data in each of the eight cell lines other than IMR-90. **(C)** Spearman's rank correlation coefficients (ρ) of enrichment scores derived from DNA methylation microarray and RRBS data in six cell lines. Cell line names are shown below the bar plots. Note that there are multiple correlation coefficients for biological or technical replicates of RRBS data in each of the six cell lines. RRBS: Reduced representation bisulfite sequencing; WGBS: Whole-genome bisulfite sequencing.

other methods estimate gene expression using all CpG-site information with linear regression [11]; and others still directly rank genes according to the overall DNA methylation level by adapting global tests [12]. Although many of these methods have seemingly reduced the bias, the following fundamental problems have remained unsolved for gene-level summarization. First, the impact of DNA methylation depends on the gene structure. For example, hypermethylation in the promoter region is associated with down-regulation in expression, whereas hypermethylation in the gene body is associated with up-regulation [1,2]. Therefore, similar DNA methylation levels may impart different consequences due to different gene structures. Adding further complexity is that gene structure can fluctuate by cell type and condition [21]. Second, a substantial number of CpG sites are located in the intergenic region (i.e., 277,806 among total 866,836 probes for the canonical transcripts [20] on Infinium MethylationEPIC array), and therefore are not easily assigned to a gene. However, such intergenic regions frequently include distal enhancers and chromatin regulatory elements that play critical roles in gene expression and are often regulated by DNA methylation [1–3]. As such, gene-level summarization of the DNA methylome remains challenging, and the use of a simplified model may confound biological interpretation.

Given the tight link between DNA methylation and TF binding [1,2], we reasoned that region-set analysis could better model the molecular events that engender the biological state. Thus far, three other informatics tools to our knowledge – LOLA [17], MIRA [16] and COCOA [18] – have been developed for region-set analysis of the DNA methylome. Whereas COCOA only calculates phenotype-correlated region-set statistics [18], LOLA statistically

evaluates the enrichment of a genomic region in a group-wise manner, and MIRA outputs a score for a sample per region set, which is most similar to Methyl-ssRSEA, the tool developed in the present study. Methyl-ssRSEA provides sample-wise analysis and generates an enrichment score for a region set in a sample, which allows for further statistical analyses, such as pairwise comparison, correlation, and clustering. Although MIRA provides similar sample-wise scores, which are tightly correlated with Methyl-ssRSEA scores in our evaluation, MIRA uses off-region methylation signals surrounding the region set and does not refer to genome-wide information for background methylation status subtraction, which may explain the slightly decreased performance of MIRA relative to Methyl-ssRSEA.

In the two biological settings, we show that region-set-level summarization efficiently and reproducibly captures hypo- or hyper-methylation in the TF-BSs known to be active or inactive in the cell type. Specifically, Methyl-ssRSEA revealed hypomethylation of OCI-Ly7 BCL6-, Jurkat MYB- and HL60 LYL1-BSs in the DNA methylomes of peripheral B, T and myeloid cells, respectively. Furthermore, ESR1-, GATA3- and FOXA1-BSs were detected as enriched in the hypomethylated regions of hormone receptor-positive breast cancers. These findings are consistent with previous observations [3,27] and validate the Methyl-ssRSEA method. It is also worth noting that Methyl-ssRSEA can deal with both microarray and sequenced-based DNA methylome data, with reproducible results measured across the platforms.

One of the advantages of the Methyl-ssRSEA tool is that we can expand its use by expanding the definition of a region set to a more complex form; for example, the genomic regions bound by two or more TFs. Starting with two region sets, e.g., ESR1-BSs and GATA3-BSs in the T47D breast cancer cell line, a new region set file could be generated by selecting ESR1-BSs that overlap with GATA3-BSs. The Methyl-ssRSEA scores could then be computed for ESR1-BSs that are also bound by GATA3 simultaneously. Furthermore, in the current study, we used ChIP-seq data for TF-BSs or histone marks as a source of functional genomic regions. However, the source can be expanded to the other modalities such as open chromatin regions derived from Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) or DNase hypersensitivity sites sequencing (DNase-seq) as well as theoretical regions, such as genomic regions with a certain TF binding motif.

Conclusion

Methyl-ssRSEA relies on a genomic region set, but not on a gene set. Integration of the function of gene set analysis with Methyl-ssRSEA may provide deeper insight into the biology of the DNA methylome in a sample. This will be important to tackle in future studies. Taking advantage of the growing number of DNA methylomes and genomic region-set data in public data repositories, Methyl-ssRSEA can be used to identify crucial TFs or genomic regions – even those not previously anticipated as having an association – in the cell with a condition or phenotype of interest. Methyl-ssRSEA thus offers a novel way to functionally interpret the DNA methylome.

Future perspective

Selection of differentially methylated regions prior to the process of Methyl-ssRSEA may be a potential scope for future study. Sample-wise region-set enrichment analysis is applicable to conventional DNA methylome data measuring 5-methylcytosine marks as well as data measuring other nucleotide modifications that can be measurable via long-read sequencing technologies, such as 5-hydroxymethylcytosine and 6-methyladenine. In addition, as discussed above, the uses of the Methyl-ssRSEA tool can be expanded by expanding the definition of a region set to a more complex form, and by expanding the source to include data from other modalities such as ATAC-seq and DNase-seq data. Furthermore, integrating the function of gene set analysis with Methyl-ssRSEA may provide further biological clues to unravel the DNA methylome. These applications will lead to a deeper understanding of the epigenetic regulation of processes including cell differentiation and cancer development.

Summary points

- Gene set analysis has commonly been used to interpret DNA methylome data. However, summarizing the DNA methylation level of a gene is challenging due to the variable number and density of CpG sites, CpG methylation levels and intergenic CpGs.
- We propose to move beyond gene set analysis by using region sets to annotate the DNA methylome. For this purpose, we developed an R-based informatics tool, single sample region set enrichment analysis for DNA methylome (Methyl-ssRSEA), to conduct sample-wise, region-set enrichment analysis.
- Methyl-ssRSEA is run on a set of genome regions representative of a cell state with altered methylation patterns, test data comprising a matrix of β values with CpG probe IDs for samples and CpG coordinates to match probe ID and genomic coordinates.
- Proof-of-concept computational experiments demonstrate that Methyl-ssRSEA recognizes the methylation profiles of peripheral blood cell types and breast cancer subtypes.
- Methyl-ssRSEA was more accurate in discriminating blood cell types when compared with other tools, such as Methylation-based Inference of Regulatory Activity and Locus Overlap Analysis, previously developed for region-set analysis.
- Methyl-ssRSEA can handle both DNA methylation microarray, and whole-genome or reduced representation bisulfite sequencing platforms, generating reproducible results.
- We conclude that Methyl-ssRSEA is useful for biological interpretation of the DNA methylome.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/epi-2021-0065

Acknowledgments

The authors thank K Inaki, Y Uemura, N Fukui, M Araki, T Hagio, T Kaneyasu, S Konno, N Matsumura, K Kiyotani, Y Nakamura, Y Miki and T Noda for helpful discussions. The authors also thank Minako Hoshida for administrative assistance and R Jackson for editing a draft of this manuscript.

Financial & competing interests disclosure

This work was supported by JSPS KAKENHI; grant numbers JP20K09634 (O Gotoh), JP18K07338 (S Mori), JP17K18337 (O Gotoh) and JP15K06861 (S Mori), by the Vehicle Racing Commemorative Foundation; grant numbers 5144, 5274 and 5393 (S Mori), and by Princess Takamatsu Cancer Research Fund; grant number 11-24317 (S Mori). JT Chang was funded by grant RP170668 from the Cancer Prevention and Research Institute of Texas. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained institutional review board approval from Japanese Foundation for Cancer Research for the research described. Since the current work is to develop a methodological framework and employed publicly available data, there was no need to obtain informed consent from the patients.

Data sharing statement

The source code of Methyl-ssRSEA.R and the example data will be available from the GitHub repository (<https://github.com/jfcr-genome/Methyl-ssRSEA>) after acceptance of the manuscript. The code for reviewer is uploaded through the manuscript submission system.

References

1. Bock C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* 13(10), 705–719 (2012).
2. Suelves M, Carrio E, Nunez-Alvarez Y, Peinado MA. DNA methylation dynamics in cellular commitment and differentiation. *Brief. Funct. Genomics* 15(6), 443–453 (2016).
3. Varley KE, Gertz J, Bowling KM *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* 23(3), 555–567 (2013).

4. Morris TJ, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* 72, 3–8 (2015).
5. Yu X, Sun S. Comparing five statistical methods of differential methylation identification using bisulfite sequencing data. *Stat. Appl. Genet. Mol. Biol.* 15(2), 173–191 (2016).
6. Muller F, Scherer M, Assenov Y *et al.* RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 20(1), 55 (2019).
7. Geleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* 29(15), 1851–1857 (2013).
8. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* 32(2), 286–288 (2016).
9. Li S, He T, Pawlikowska I, Lin T. Correcting length-bias in gene set analysis for DNA methylation data. *Statistics Interface* 10(2), 279–289 (2017).
10. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics* 35(11), 1958–1959 (2019).
11. Wang Y, Franks JM, Whitfield ML, Cheng C. BioMethyl: an R package for biological interpretation of DNA methylation data. *Bioinformatics* 35(19), 3635–3641 (2019).
12. Dong D, Tian Y, Zheng SC, Teschendorff AE. ebGSEA: an improved Gene Set Enrichment Analysis method for epigenome-wide-association studies. *Bioinformatics* 35(18), 3514–3516 (2019).
13. Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* 102(43), 15545–15550 (2005).
14. Barbie DA, Tamayo P, Boehm JS *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462(7269), 108–112 (2009).
15. Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC Bioinformatics* 19(Suppl. 5), 115 (2018).
16. Lawson JT, Tomazou EM, Bock C, Sheffield NC. MIRA: an R package for DNA methylation-based inference of regulatory activity. *Bioinformatics* 34(15), 2649–2650 (2018).
17. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 32(4), 587–589 (2016).
18. Lawson JT, Smith JP, Bekiranov S, Garrett-Bakelman FE, Sheffield NC. COCOA: coordinate covariation analysis of epigenetic heterogeneity. *Genome Biol.* 21(1), 240 (2020).
19. Sugiyama Y, Gotoh O, Fukui N *et al.* Two distinct tumorigenic processes in endometrial endometrioid adenocarcinoma. *Am. J. Pathol.* 190(1), 234–251 (2020).
20. Kaneyasu T, Mori S, Yamauchi H *et al.* Prevalence of disease-causing genes in Japanese patients with BRCA1/2-wildtype hereditary breast and ovarian cancer syndrome. *NPJ Breast Cancer* 6(25), 1–13 (2020).
21. Encode Project Consortium, Moore JE, Purcaro MJ *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583(7818), 699–710 (2020).
22. Cheneby J, Menetrier Z, Mestdagh M *et al.* ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 48(D1), D180–D188 (2020).
23. Sanchez-Castillo M, Ruau D, Wilkinson AC *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* 43(Database issue), D1117–1123 (2015).
24. Hannon E, Mansell G, Burrage J *et al.* Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology. *bioRxiv* doi:10.1101/2020.05.21.107730 (2020).
25. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418), 61–70 (2012).
26. Fortin JP, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* 33(4), 558–560 (2017).
27. Fleischer T, Tekpli X, Mathelier A *et al.* DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* 8(1), 1379 (2017).
28. Schwender H. Siggenes: Multiple Testing using SAM and Efron's Empirical Bayes Approaches. *R package version 1.60.0*. doi:10.18129/B9.bioc.siggenes (2019).
29. Zou LS, Erdos MR, Taylor DL *et al.* BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics* 19(1), 390 (2018).