Taryma-Leśniak *et al.*
Epigenetics & Chromatin (2024) 17:30
https://doi.org/10.1186/s13072-024-00555-5

Epigenetics & Chromatin

RESEARCH

Open Access

Methylation patterns at the adjacent CpG sites within enhancers are a part of cell identity

Olga Taryma-Leśniak[1], Jan Bińkowski[1], Patrycja Kamila Przybylowicz[1], Katarzyna Ewa Sokolowska[1], Konrad Borowski[1] and Tomasz Kazimierz Wojdacz[1*]

Abstract

Background It is generally accepted that methylation status of CpG sites spaced up to 50 bp apart is correlated, and accumulation of locally disordered methylation at adjacent CpG sites is involved in neoplastic transformation, acting in similar way as stochastic accumulation of mutations.

Results We used EPIC microarray data from 596 samples, representing 12 healthy tissue and cell types, as well as 572 blood cancer specimens to analyze methylation status of adjacent CpG sites across human genome, and subsequently validated our findings with NGS and Sanger sequencing. Our analysis showed that there is a subset of the adjacent CpG sites in human genome, with cytosine at one CpG site methylated and the other devoid of methyl group. These loci map to enhancers that are targeted by families of transcription factors involved in cell differentiation. Moreover, our results suggest that the methylation at these loci differ between alleles within a cell, what allows for remarkable level of heterogeneity of methylation patterns. However, different types of specialized cells acquire only one specific and stable pattern of methylation at each of these loci and that pattern is to a large extent lost during neoplastic transformation.

Conclusions We identified a substantial number of adjacent CpG loci in human genome that display remarkably stable and cell type specific methylation pattern. The methylation pattern at these loci appears to reflect different methylation of alleles in cells. Furthermore, we showed that changes of methylation status at those loci are likely to be involved in regulation of the activity of enhancers and contribute to neoplastic transformation.

Keywords Epigenomics, Epigenetics, DNA methylation, Co-methylation, Methylation patterns

Background

The covalent addition of a methyl group to cytosine, referred to as DNA methylation, plays a key role in the regulation of gene expression [1]. In humans, DNA methylation occurs almost exclusively at CpG dinucleotides, which are non-randomly distributed in the genome, with the regions of the higher-than-expected density of CpGs referred to as CpG islands (CGIs) [2]. It is generally accepted that the methylation of two consecutive CpG sites is correlated [3–8] and the same methylation status of adjacent CpG sites, referred to as co-methylation, is considered to be essential for regulatory function of CGIs (as reviewed e.g., in [9]). Consequently, following the principle of co-methylation of the adjacent CpG sites, methylation pattern differences between healthy and pathologically changed cells and tissues have predominantly been studied as; average methylation levels at single CpG sites displaying different variability of methylation between tissues [10], haploblocks consisting minimum of three highly coregulated CpG sites [5] or regions containing a number of CpG sites with corelated

*Correspondence:
Tomasz Kazimierz Wojdacz
tomasz.wojdacz@pum.edu.pl
[1] Independent Clinical Epigenetics Laboratory, Pomeranian Medical University in Szczecin, 71-252 Szczecin, Poland

methylation levels referred to as Differentially Methylated Regions (DMRs) [11–14].

The adjacent CpG sites that do not follow principle of co-methylation have been observed in neoplastic cells, but discordant methylation status of these CpG sites was attributed to "stochastically disordered methylation in malignant cells" [15], which similarly to genetic instability, provides random ability of cancer cells to search for superior evolutionary trajectories [15–19].

In our study, we identified a subset of adjacent CpG sites in methylomes of healthy human cells, that do not follow the principle of co-methylation and display cell type specific discordant methylation patterns. We further show that the intercellular heterogeneity of methylation patterns at these loci may result from different methylation of adjacent CpG site at specific alleles in the cell. That, can potentially allow for regulation of binding of proteins targeting these loci such as Methyl-CpG-Binding Domain proteins. Moreover, our results indicate that these CpG sites do not map to the borders of deferentially methylated regions but predominantly to enhancers which are targeted by families of transcription factors involved in cell differentiation. Finally, we show that cell type specific fingerprints of methylation at these loci are lost during malignant transformation.

Methods

Data collection

Our study utilized 112 EPIC microarray methylation profiles from four independent studies: GSE123914 (this data set included arrays for 35 individual blood samples, 34 with DNA methylation measurements at two timepoints approximately one year apart) [20], GSE153211 (n = 8 arrays) [21], GSE112618 (n = 6 arrays) [22] and GSE166844 (n = 29 arrays; from 29 individuals, with 14 pairs being monozygotic twins) [10].

The analysis of cell type specific discordant methylation pattern was performed using EPIC microarrays obtained for purified white blood cell fractions, including: GSE110554 (neutrophils (n = 6), monocytes (n = 6), B cells (n = 6), CD4 + T cells (n = 7, six individual arrays and one technical replicate), CD8 + T cells (n = 6), NK cells (n = 6)) [22], and GSE166844 (granulocytes (n = 29), monocytes (n = 28), B cells (n = 28), CD4 + T cells (n = 28), CD8 + T cells (n = 28) [10].

Changes of discordant methylation during neoplastic transformation were investigated using EPIC methylation profiles of blood samples from patients with acute myeloid leukemia (AML; n = 458) from dataset GSE124413 [23] and with chronic lymphocytic leukemia (CLL; n = 114) previously described in [24]. The discordant methylation in healthy tissues was assessed using: GSE121377 (n = 7) dataset for thyroid [25], GSE124413 (n = 41) for bone marrow [23], GSE142141 (n = 47) for skeletal muscle, GSE100850 (n = 5) for breast tissue [26], and GSE132804 (n = 206) for colon tissue [27].

Data pre-processing

Raw Illumina MethylationEPIC array data (.idat) were processed, QC (Quality Control) checked, and normalized using Beta Mixture Quantile (BMIQ) method [28] in the ChAMP Package (R/Bioconductor) [29]. All the single nucleotide polymorphisms (SNPs) shown to influence methylation analysis results were filtered out using ChAMP pipeline [30].

Identification of discordantly methylated adjacent CpG loci

We used a linear regression model based on ordinary least square method to analyze methylation status of adjacent CpG loci (across panels of samples), defined as two CpG sites targeted by EPIC microarray, spaced less than 50 bp apart and with no additional CpG sites, not assayed by the microarray, in between. We named CpG sites within adjacent CpG loci as base and consecutive CpG site. For each of these loci, we estimated linear model defined as follows: methylation level ~ Intercept + CpG position, where CpG position is equal to 0 if CpG is base and 1 if CpG is consecutive in the genomic context. A locus was considered discordantly methylated, if |slope| coefficient (reflecting difference between methylation levels of base and consecutive CpGs, within adjacent CpG loci) in the model, was more than 0.3 and FDR corrected p-value (t-test) for this coefficient was ≤ 0.05, while the loci with |slope| less than 0.05 or FDR corrected p-value > 0.05 CpG was considered co-methylated. Importantly, all adjacent CpG sites with an additional CpG site between consecutive and base CpG, not assayed by EPIC microarray, were removed from the analysis. The Methrix R package was used to extract annotations of all CpGs in hg19 reference genome.

High-dimensional data visualization

To visualize methylation differences at the adjacent CpG loci, we calculated delta-beta values, which are the difference between beta-values of the base and consecutive CpG site within locus. The delta-beta values are in the range from − 1 to 1, where values close to 0 indicate co-methylation, and values close to 1 and − 1 indicate discordant methylation between base and consecutive CpG site within a locus, depending on the direction of the analysis. Data with more than three dimensions were visualized in the form of heatmaps, clustered using unsupervised Ward's algorithm and standardized Euclidean distance.

Enrichment analysis for genomic regions

Genomic context enrichment analysis was performed using Genomic Locus Overlap Enrichment Analysis (LOLA) package [31] and two custom databases, generated as described in package documentation. The first database was generated using "Relation_to_UCSC_CpG_Island" and "UCSC_RefGene_Group" annotations from Infinium Methylation EPIC BeadChip manifest (Illumina). The second database was generated using 15 types of genomic segments predicted based on five histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) for white blood cells (samples id: E034, E062, E044, E043, E047, E048, E029, E035, E032, E046, E030), downloaded from the [32]. To test the statistical significance of enrichment, we used the default one-sided Fisher exact test. Analysis of transcription factor (TF) motif enrichment was performed using the findMotifsGenome.pl script from HOMER [33], with the following HOMER specific parameters: hg19 reference genome, mask repeats/lower case sequence, CpG normalization, and with all assessed EPIC adjacent CpG sites as a background. The motif enrichment was determined using hypergeometric test.

Validation of discordant methylation patterns

We validated the methylation patterns identified at adjacent CpG sites in blood cell types using deep WGBS data, generated for: granulocytes (n = 3), monocytes (n = 3), B cells (n = 3), CD4T cells (n = 3), CD8T (n = 3), NK cells (n = 3) (GSE186458) [14]. We also used Sanger sequencing analysis to validate methylation patterns at representative adjacent CpG loci in blood. Here, peripheral blood samples were collected from 24 healthy individuals, under individuals' informed consent. The study was approved by Ethics Committee of the Pomeranian Medical University in Szczecin (within the nr KB-0012/56/2021). Genomic DNA was extracted from peripheral blood samples of 24 patients, using salting out method, as described previously in: [34]. The quantity and quality of obtained DNA were assessed using the Qubit™ dsDNA BR Assay and Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Bisulfite conversion of DNA (500 ng) was carried out using the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA), according to the manufacturer's protocol. Bisulfite treated DNA (v = 3 µl) were amplified in 25 µl reaction volume containing: PCR Master Mix (2×) (Thermo Fisher Scientific, Waltham, MA, USA), 4uM forward and reverse primers, and with the following protocol: 94 °C for 5 min, 38 × [94 °C for 30 s, 55–60 °C for 30 s, 72 °C for 1 min], 72 °C for 8 min. The length of the PCR product was tested on 2% agarose gel, and sent for Sanger sequencing to Genomed S.A. (Warsaw, Poland).

Quantification and statistical analysis

Python version 3.9.7 and R version 4.1.2 were used for statistical analyses. All statistical analyses were corrected for multiple comparison, using the Benjamini–Hochberg method. The statistical significance level assumed in this study was equal to 0.05.

Results

A subset of adjacent CpG sites in the human genome does not follow the principle of co-methylation

Identical methylation status of adjacent CpG sites (methylated or not methylated), referred to as co-methylation, is considered essential for the regulatory function of CGIs [9]. We analyzed genome-wide methylation dynamics between adjacent CpG sites at 87 258 CpG site pairs present on the EPIC microarray, spaced less than 50 bp apart and with no additional CpG sites not assayed by the microarray in between. We named these CpG pairs "adjacent CpG loci" and calculated methylation level difference (delta beta-value, see Methods for details) between base and the consecutive CpG site, constituting adjacent CpG loci.

As expected, initial analysis of methylation at adjacent CpG loci in 112 EPIC microarrays from four different studies obtained for peripheral blood samples [10, 20–22], showed, that vast majority of CpG sites within adjacent CpG loci are co-methylated and display identical methylation status (delta beta-value close to 0; Fig. 1a—white color in the heatmap).

Interestingly, however, we also identified a subset of adjacent CpG loci, within which base and consecutive CpG site had different statuses of methylation in all analyzed samples and named these, discordantly methylated (Fig. 1a, red or blue color in the heatmap). We then performed identical analysis using methylation profiling data from 306 microarrays from five different types of healthy tissues: bone marrow, thyroid, breast, colon and skeletal muscle [23, 25–27]. Similarly, to the previous analysis, a subset of adjacent CpG sites was discordantly methylated in each of the tissue types (Fig. 1b), but more importantly some of these adjacent CpG loci were co-methylated in one tissue type and discordantly methylated in others, suggesting tissue specificity of methylation patterns at the adjacent CpG loci.

Developmentally close cells and tissues display similar discordant methylation patterns at a subset of adjacent CpG sites

Measurement of methylation levels with microarray technology always represents average methylation level of all cells in the sample, and this limits analyses of cell specific methylation patterns (as reviewed in: [35]). Therefore, to elaborate the significance of changes of methylation at

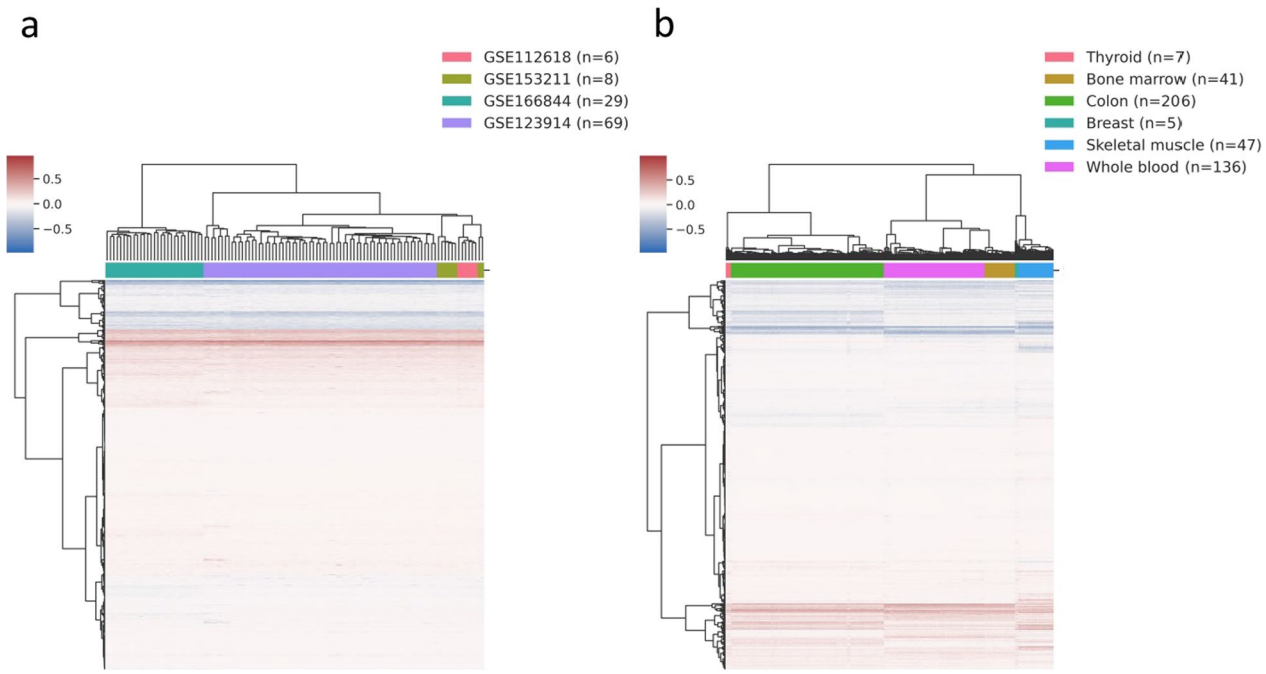

Fig. 1 Methylation patterns at adjacent CpG loci in different types of healthy tissues. Unsupervised hierarchical clustering of samples based on methylation differences (delta beta-value) between base and consecutive CpG site of adjacent CpG loci, **a** in healthy blood samples from four independent studies and **b** in different types of healthy tissues, including thyroid, bone marrow, colon, breast, skeletal muscle, and whole blood

adjacent CpG loci of specific cell type, we analyzed EPIC microarrays data obtained for six fractions of sorted white blood cells, including granulocytes, monocytes, B cells, CD4 + T cells, CD8 + T cells, and natural killer (NK) [10, 22].

We measured the methylation difference between base and consecutive dinucleotide within adjacent CpG loci using regression model and considered an adjacent CpG loci as discordantly methylated when that difference was more than 30 percentage points (pp) ($|slope| \geq 0.3$ and FDR-corrected p value ≤ 0.05, t-test). With technical limitations of the BeadChip array technology in mind, that methylation difference is likely to reflect presence of methylation at one of the alleles of the adjacent CpG loci in the cell.

This analysis identified different numbers of discordantly methylated adjacent CpG loci in each of the cell types, specifically 2049 in granulocytes, 1905 in monocytes, 1909 in B cells, 2066 in CD4 + T cells, 1986 in CD8 + T-cells, and 1927 in NK cells (results of the analysis for each cell type are shown in Supplementary Tables 1–6). Unsupervised clustering, based on delta beta-values of all identified adjacent loci showed a very specific clustering of the samples by cell type, with the majority of the loci displaying changes of methylation status between discordantly methylated and co-methylated in different cell types (Fig. 2a). Only CD4 + T cells and CD8 + T cells mixed together, but those cells are developmentally very close.

We then performed pairwise comparison of adjacent CpG loci displaying discordant methylation status, between types of blood cells included in our analysis. The results (Fig. 2b) showed, that CD4 + T cells and CD8 + T cells, together with granulocytes and monocytes, which are the most developmentally close cell lines in our analysis, had the largest number of common discordantly methylated adjacent CpG loci, 0.81 and 0.71, respectively. Whereas, for example, granulocytes and CD8 + T cells, together with monocytes and CD8 + T, which represent the most developmentally distant cells in our analysis, had the lowest number of common adjacent CpG loci with discordant methylation status, 0.40 and 0.41, respectively (Fig. 2b). These results suggested the function of discordant methylation at adjacent CpG loci in cell specialization.

We also compared the number of adjacent CpG loci with discordant methylation status between thyroid, bone marrow, colon, breast, skeletal muscle and whole blood (Supplementary Fig. 1). Again, the number of common adjacent CpG loci was the highest for developmentally close tissues, such as blood and bone marrow (0.71) and the lowest for developmentally distant tissues such as breast and bone marrow (0.21).

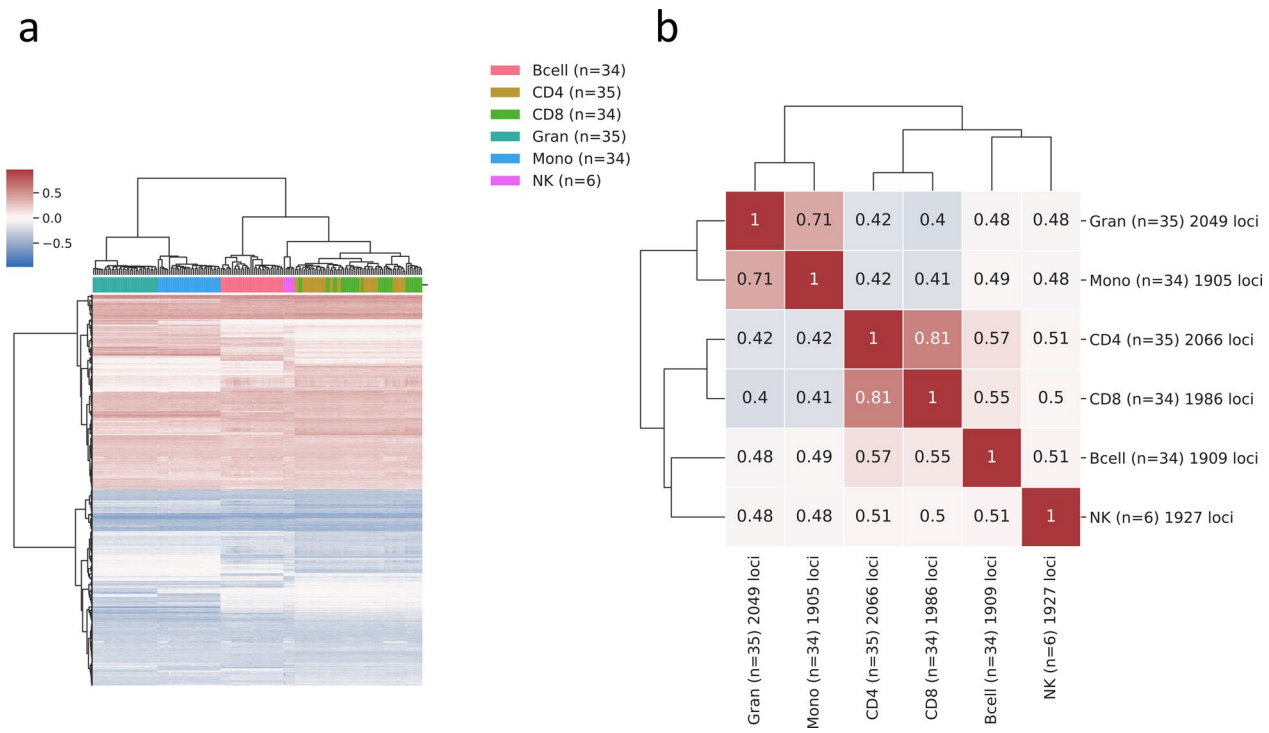

Fig. 2 Methylation patterns at adjacent CpG loci with discordant methylation status in six types of sorted white blood cells. **a** Unsupervised hierarchical clustering of white blood cells (WBC) based on delta beta-values of CpG pairs identified in different cell types. **b** Pairwise comparison of identified loci between each two types of WBC. Boxes represent the Jaccard similarity index between the cell types

A subset of adjacent CpG loci displays highly heterogeneous methylation patterns

Next, we went on to analyze methylation patterns of base and consecutive CpG sites within adjacent CpG loci. Not surprisingly and in line with the definition of CpG island, the base and the consecutive CpG site at co-methylated adjacent CpG loci (n = 40 404) (|slope| ≤ 0.05 and FDR-corrected p value ≤ 0.05, t-test test) were predominantly either methylated or not methylated in all analyzed cell types (Fig. 3a, b). Several of the co-methylated CpG loci in this analysis (n = 81; 0.2%), displayed about 50% (beta-value ≤ 0.6 and ≥ 0.4) methylation levels in both CpG sites (Fig. 3c). It is important to mention here that, this level of methylation recorded with BeadChip array can be attributed to the uniform (50–50) mixture of cells sub-populations, with both alleles methylated and not methylated in those cells, or one allele methylated and other non-methylated within one cell (Fig. 3c, right panel). However, because in our analysis, we used data from relatively pure cell populations, and moreover, we did not observe high levels of methylation measurement variance at analyzed adjacent CpG loci, it is most likely that these methylation patterns are attributed to methylation of both CpG sites, on only one of the alleles, possibly consequence of the phenomenon similar to genomic imprinting [36]. Also, a subset of 308 of the co-methylated loci in this analysis, displayed more than 30 percentage points (pp) of methylation level difference in at least one of the analyzed cell types (Fig. 3d). These loci may mark genes monoallelically expressed in different cell types [37] and potentially regulate allele specific binding of proteins such as Methyl-CpG-Binding Domain proteins [38].

Then, we analyzed methylation patterns within adjacent CpG loci with discordantly methylated base and consecutive CpG sites (n = 1890; |slope| ≥ 0.3 and FDR-corrected p value ≤ 0.05, t-test) in at least one of the analyzed cells. About half of those loci (n = 850) displayed identical methylation difference between base and consecutive CpG site in all analyzed cell types, and the remaining loci (n = 1040) were discordantly methylated in at least one of the analyzed cells. Again, considering that in general, the methylation level measurements in our analysis displayed a very low level of variance between different cell types, the methylation levels observed at this subset of adjacent CpG loci, may reflect one of the eight variants of simultaneous co-methylation of one and discordant methylation of the other allele illustrated in (Fig. 4a). Interestingly, in this subset of adjacent CpG loci, there were a few loci with almost 100 pp methylation difference between base and consecutive CpG site (five in

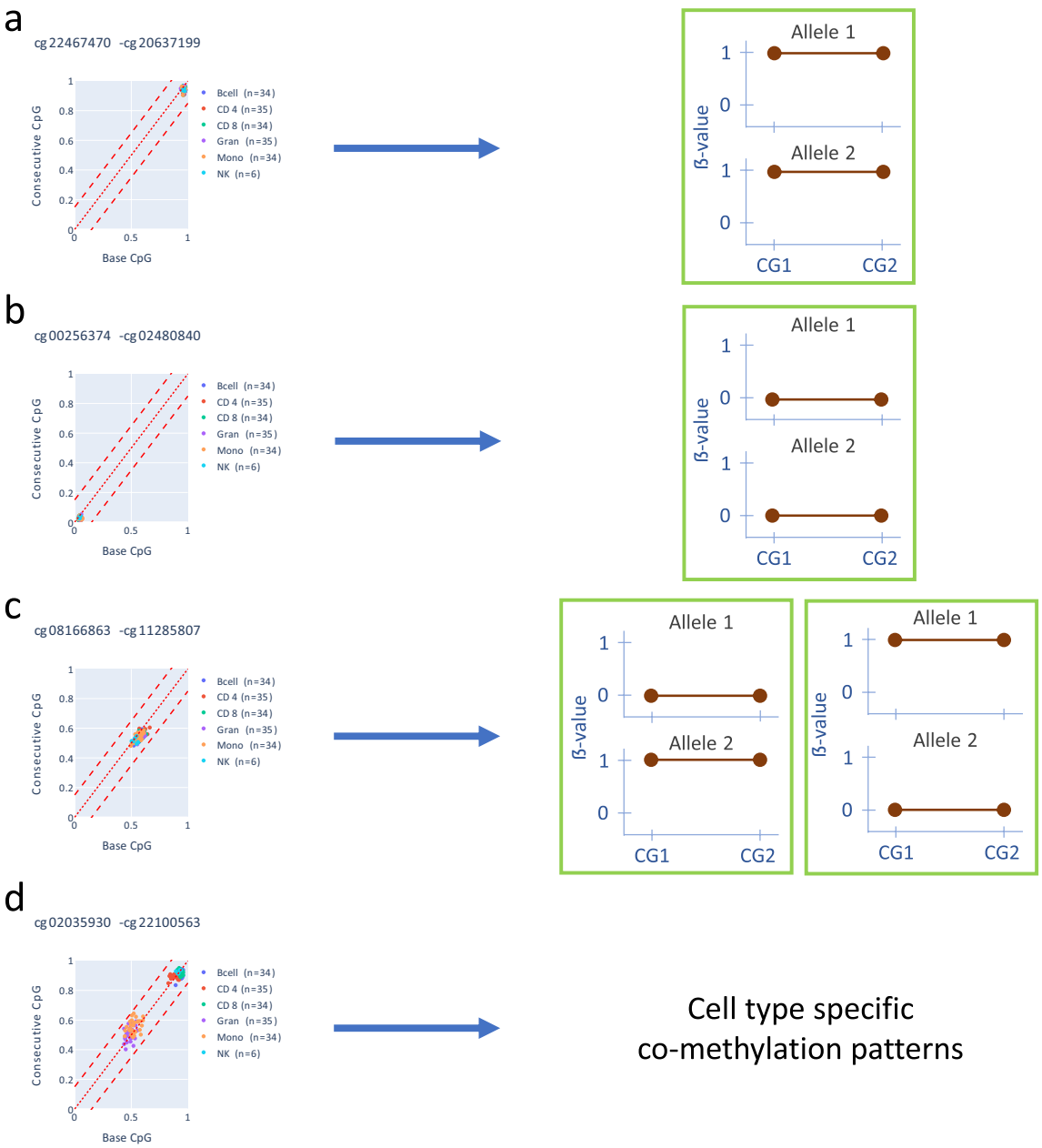

Fig. 3 Types of methylation patterns within adjacent CpG loci displaying identical methylation (co-methylation) at base and consecutive CpG site of adjacent CpG loci. Plots (left panels) illustrate the association of methylation levels between base and consecutive CpG site, with the solid line indicating no methylation level difference and dashed lines ± 0.15 confidence interval, and graphical illustration (right panels) of those methylation levels at single cell level, **a** loci with base and consecutive CpG sites methylated, **b** loci with both base and consecutive CpG site not methylated, **c** loci with 50% methylation level at both base and consecutive CpG site, and **d** loci displaying different types of co-methylation patterns between the cell types, with 50% methylation level in granulocytes and monocytes, as well as 100% methylation levels in B cells, NK, CD4T and CD8T

B cell, five in CD4, five in CD8, 13 in Gran, 13 in Mono, and nine in NK; with $|slope| \geq 0.80$, FDR < 0.05) (Fig. 4b). This confirms that base and consecutive CpG site at adjacent CpG loci can acquire a different methylation status at each of the alleles in a cell because observed at those loci methylation levels are unlikely to be explained by the mixture of cell subpopulations. Also, a number of discordantly methylated loci had different methylation patterns in specific cell types (Fig. 4c). Overall, these results illustrate the vast heterogeneity of methylation patterns

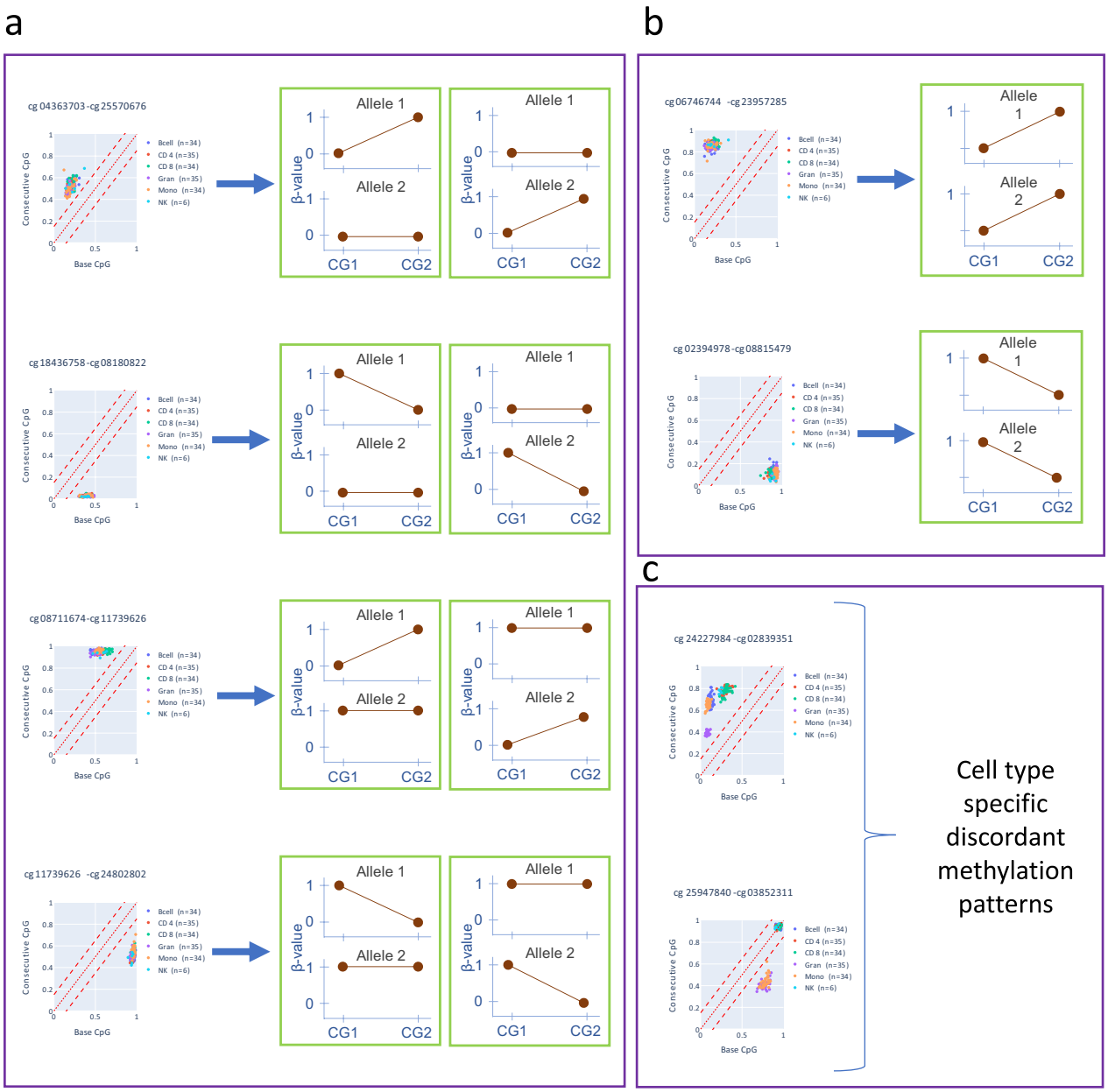

Fig. 4 Types of methylation patterns within adjacent CpG loci with discordantly methylated base and consecutive CpG site. **a**–**c** Plots (left panels) illustrate the association of methylation levels between base and consecutive CpG site, with the solid line indicating no methylation level difference and dashed lines ± 0.15 confidence interval, as well as graphical illustration (right) of those methylation levels at single cell level, in **a** loci with about 50 pp methylation level difference between base and consecutive CpG site in all analyzed cell types, **b** loci with almost 100 pp methylation level difference between base and consecutive CpG site in all analyzed cell types, and **c** loci with discordant methylation patterns specific for the cell type

at adjacent CpG sites and remarkable stability of one of those patterns in the specific cell type.

Different patterns of methylation at adjacent CpG loci mark specific regulatory regions of the genome

To understand if discordant methylation patterns are associated with particular biological functions, we subdivided analyzed adjacent CpG loci into four subsets: "common co-methylated loci" (n = 40,404) (Fig. 5a), displaying similar patterns of co-methylation in all cell types included in our analysis; "co-methylated cell specific loci" (n = 308) (Fig. 5b), displaying different co-methylation patterns between the cell types (these loci are in principle differentially methylated regions—DMRs, if we define

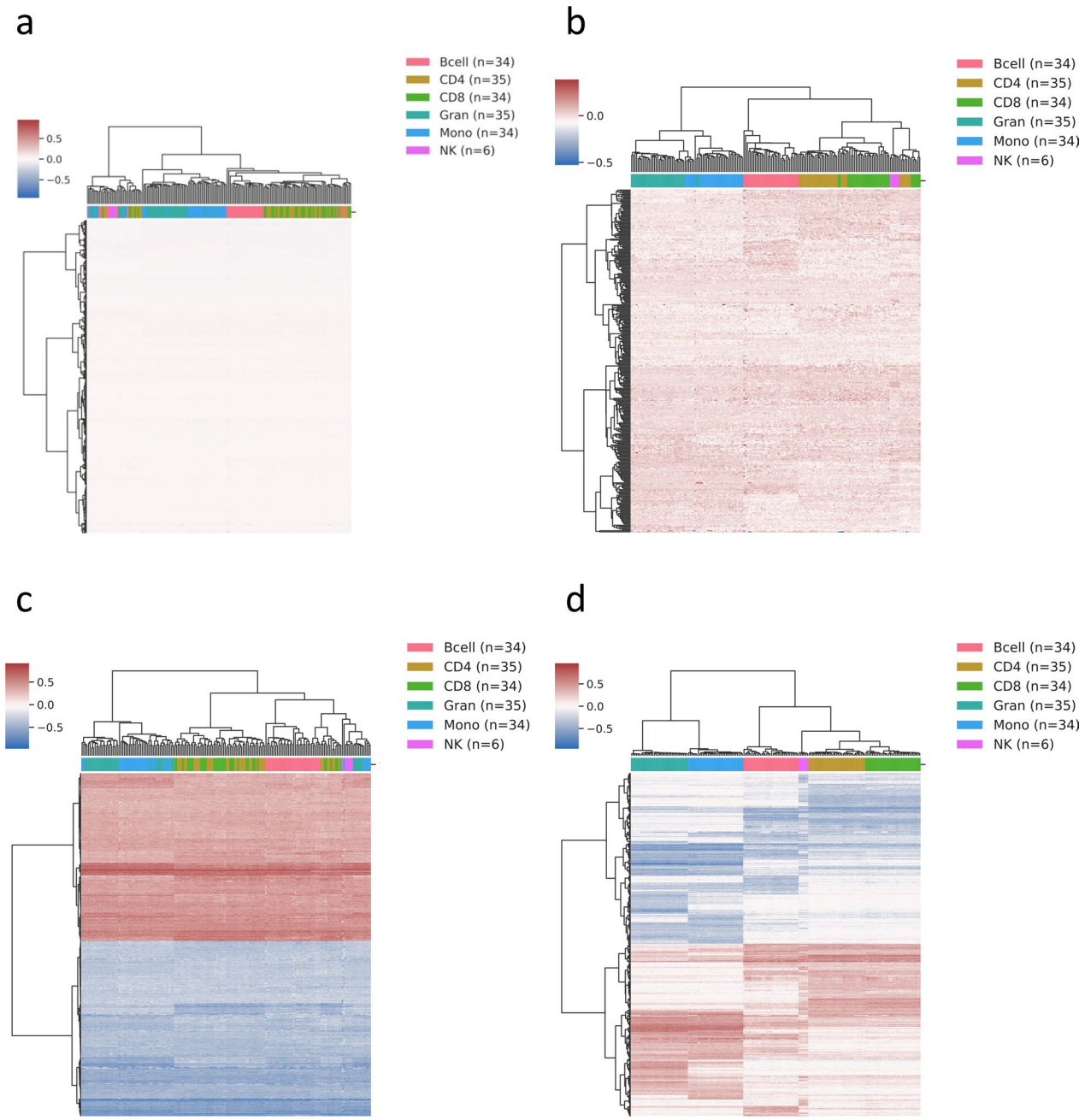

Fig. 5 Types of adjacent CpG loci in six types of sorted white blood cells. Results of unsupervised hierarchical clustering of six types of sorted white blood cells, based on delta beta-values at four subsets of discordantly methylated adjacent loci, including, **a** "common co-methylated", **b** "co-methylated cell specific", **c** "common discordantly methylated", and **d** "discordantly methylated cell specific" adjacent CpG loci

that region as a region with two consecutive CpG sites with the identical methylation status); "common discordantly methylated loci" (n = 850) (Fig. 5c), displaying significant difference in methylation level between base and consecutive CpG site in all analyzed cell types; and "discordantly methylated cell specific loci" (n = 1040) (Fig. 5d), displaying more than 0.3 difference in delta beta-value between at least two types of blood cell types included in our analysis.

Each type of methylation patterns that we identified could potentially play different role in cell physiology. Therefore, we analyzed association of different subsets of adjacent CpG loci with specific histone marks and regulatory regions of the genome, as well

as performed gene set enrichment analysis (GSEA) of genes annotated to each of the categories of the loci. We also assessed enrichment of each type of adjacent CpG sites within transcription factors (TFs) binding sites.

The association of adjacent CpG loci displaying different patterns of methylation with histone marks in 11 types of cells including hematopoietic stem cells, from Roadmap Epigenomics Core 15-state Model [32], was analysed using Locus Overlap Analysis (LOLA) [31] (Fig. 6a–d). All CpG spaced in up to 50 bp targeted by EPIC microarray were used as background in these analyses. The results showed significant enrichment (FDR corrected p-value ≤ 0.05, Fisher exact test; odds ratio (OR) > 2) of "common co-methylated" loci in most of the analyzed cell types in regions marked with histones containing modifications associated with active transcription states (TssA) and bivalent regulatory states (TssBiv) (Fig. 6a), not surprisingly indicating that identical methylation status of the adjacent CpG sites within CGI is essential for the regulatory function of those genomic regions. The "co-methylated cell specific" adjacent CpG loci, were associated with histone marks, characteristic for transcribed state at the 5′ and 3′ end of genes that show both promoter and enhancer signatures (TxFlnk), and two of the enhancer related histone marks: enhancers (Enh), and genic enhancers (EnhG) (Fig. 6b). The "common discordantly methylated" loci were associated with histones occupying three repressive chromatin states including constitutive heterochromatin (Het) and repressed Polycomb states (ReprPC, ReprPCWk), as well as one of the enhancers related states (Enh) (Fig. 6c). The "discordantly methylated cell specific" loci, which displayed most heterogeneous methylation between the cell types showed significant enrichment in the same regions as "co-methylated cell specific" loci (Fig. 6d).

Also using LOLA platform, we analyzed distribution of adjacent CpG loci in relation to CGI. As expected, "common co-methylated" adjacent CpG loci were significantly enriched in CGIs (OR = 4.85). The "co-methylated cell specific" in S-Shelf (OR = 2.65) and OpenSea (OR = 3.67). The "common discordantly methylated" adjacent CpG loci were enriched in N-Shore (OR = 2.19), and S-Shore (OR = 2.17). The "discordantly methylated cell specific" loci were enriched in N-Shelf (OR = 2.46), S-Shelf (OR = 2.14) and OpenSea (OR = 3.82) (Fig. 6e). This indicates that discordantly methylated adjacent CpG sites are not only an attribute of bordering regions of CGIs. Overall, these results show that adjacent CpG sites with most dynamic changes of DNA methylation levels in blood cells mark gene enhancers, and are located in OpenSea compartment of the chromatin.

The GSEA were based on genes annotated to each subset of the adjacent CpG loci according to relevant microarray manifest and performed using GENE2FUNC function of FUMA GWAS (Functional Mapping and Annotation of Genome-Wide Association Studies) [39] with Molecular Signatures Database (MSigDB) [40, 41] used as reference. This analysis showed, that genes annotated to "common co-methylated cell specific" and "discordantly methylated cell specific" adjacent CpG loci displaying cell type specific methylation, were enriched in GO biological categories related to specialized functions of blood cells, such as cell activation, defense response, T-cell activation or adaptive immune system (Supplementary Table 7, 8). At the same time, we did not observe significant enrichment of genes annotated to "common co-methylated" and "common discordantly methylated" loci in any of the interrogated gene set categories. However, this is not surprising because those loci predominantly constitute CGIs and majority of human genes contains CGIs in promoters. Overall, these results suggests that adjacent CpG loci displaying heterogenous methylation between specialized cells, mark genes involved in specific cellular function, and patterns of methylation of those loci may reflect regulation state of those genes in specific cells. Contrary to the genes marked by "common co-methylated" and "common discordantly methylated" loci, which appear to mark genes involved in general cell physiology, expression of which is similar in all analyzed cells.

Finally, we used HOMER platform to assess enrichment of specific categories of adjacent CpG loci within transcription factors binding sites [33]. This analysis showed that "common co-methylated" loci are significantly enriched (q-value ≤ 0.05, hypergeometric test) in regions with 113 TF binding motifs (Supplementary Table 9), what is consistent with the results showing that those loci constitute CGIs that annotate to genes involved in general cell physiology. None of the specific TF binding site motifs was found to be enriched in regions with "common discordantly methylated" loci, but chromatin state association analysis showed that this type of adjacent CpG loci, are unlike any of the other analyzed loci, associated with repressed regions of chromatin. The TF binding motifs enriched in regions harboring "co-methylated cell specific" (n = 10, Supplementary Table 10) and "discordantly methylated cell specific" (n = 48, Supplementary Table 11) adjacent CpG loci were very similar and predominantly different from TF binding motifs annotated to "common co-methylated" adjacent CpG loci (Fig. 6f and Supplementary Table 12). The analysis of function of TFs binding motifs annotated to "co-methylated cell specific" and "discordantly methylated cell specific" loci, showed that these TFs are involved hematopoietic cells development and differentiation. This analysis has a limited power but suggests that adjacent CpG loci displaying

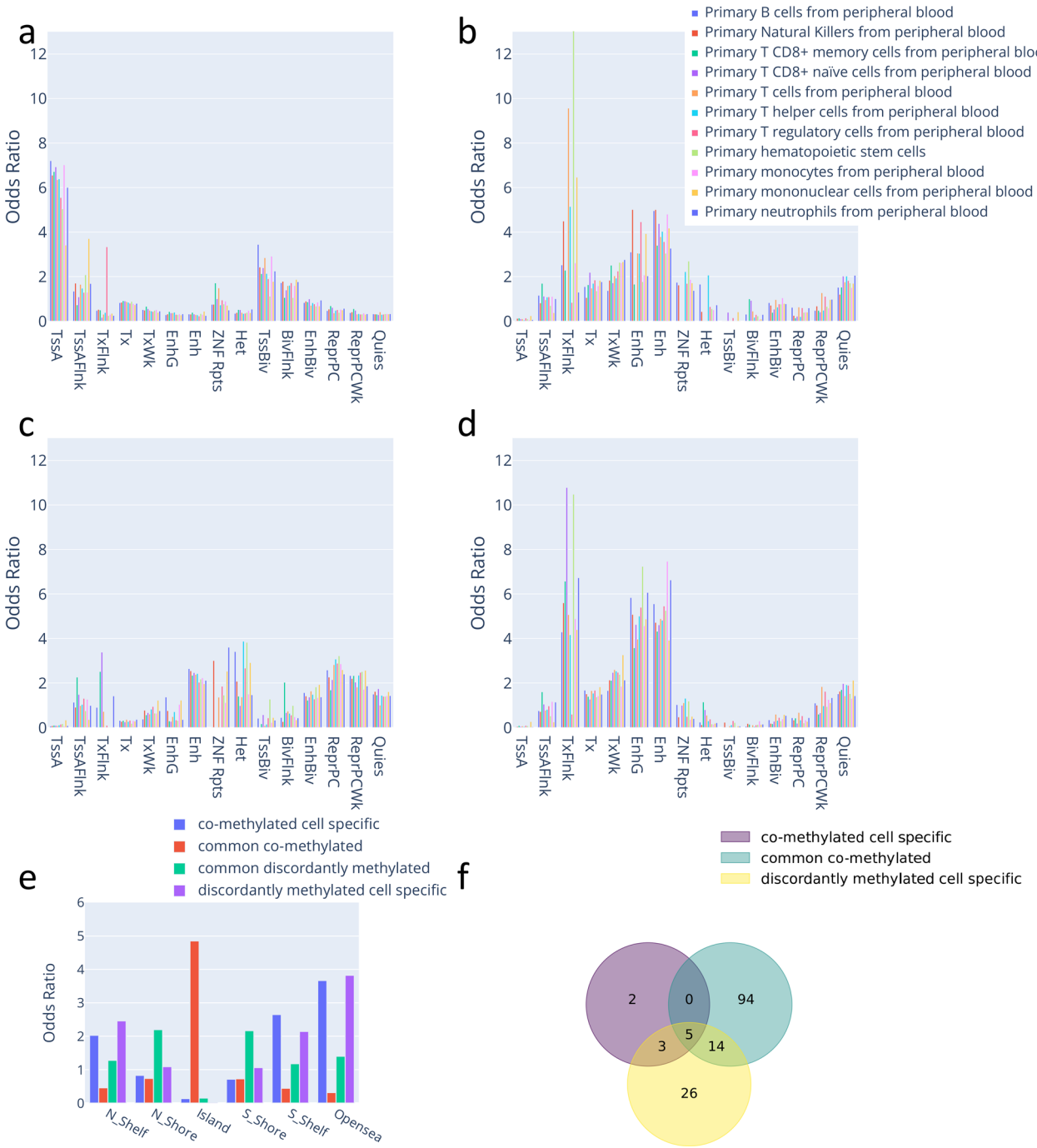

Fig. 6 Illustration of the genomic distribution of "common co-methylated", "co-methylated cell specific", "common discordantly methylated", and "discordantly methylated cell specific" loci. **a**–**d** LOLA-based, region set enrichment analysis of the regions with **a** "common co-methylated", **b** "co-methylated cell specific", **c** "common discordantly methylated", and **d** "discordantly methylated cell specific" adjacent loci analyzed loci in specific WBC types, using Core 15-state model. **e** Results of enrichment analysis of each type of adjacent CpG loci in relation to CpG island elements . **f** Venn diagram illustrating the overlap between transcription factor binding motifs identified for different subsets of adjacent CpG loci.

cell type specific methylation patterns mark TF binding motifs that bind transcription factors involved in specialized cell functions. That corroborates the results

indicating that those loci also annotate to genes involved in cell function specialization.

Methylation at adjacent CpG loci changes during neoplastic transformation

So far, locally disordered methylation at CpG sites has been reported and considered a stochastic event in cancer development [15]. We lastly analyzed whether this phenomenon affects methylation patterns at the adjacent CpG loci that display stable discordant methylation in all WBC types (Fig. 5c) during carcinogenesis of acute myeloid leukemia (AML) [23] and chronic lymphocytic leukemia (CLL) [24]. We found that the methylation patterns at those loci undergo general deregulation that appears to be stochastic and involves majority of those loci (Fig. 7a and b). Interestingly however, we did identify loci that appeared to maintained methylation pattern observed in healthy cells throughout malignant transformation (Fig. 7c), as well as loci that changed methylation status specifically in one neoplasia (Fig. 7d).

We also compared overall standard deviation (SD) of the beta-values at discordantly methylated CpG sites between CLL, AML and healthy blood. The violin plots in Fig. 7e show that the SD of the delta beta-values in both CLL (0.215 (95% CI 0.211–0.220)), and AML (0.150 (95% CI 0.147–0.153)) is significantly increased as opposed to the rather stable SD in healthy blood (0.058 (95% CI 0.057–0.060). Moreover, the average level of the standard variation was statistically significantly different (FDR corrected p-value ≤ 0.05, Mann–Whitney U test) between all compared groups.

Validation of discordant methylation phenomenon using NGS and Sanger sequencing

We validated our findings using currently available whole-genome bisulfite sequencing (WGBS) data obtained for the pools of FACS sorted cells [14] and Sanger sequencing. Identical to analysis based on EPIC microarray data, using NGS data we were able to unambiguously classify each type of blood type cells (Supplementary Fig. 2).

Sanger sequencing of representative adjacent CpG loci that we found to be discordantly methylated in whole blood using EPIC data (Fig. 1a) also unambiguously confirmed the discordant methylation patterns of those loci in each of the 24 whole blood samples that we used in the validation experiment. The representative Sanger sequencing chromatograms from this analysis are shown in Supplementary Fig. 3a–c, where for example first chromatogram (Supplementary Fig. 3a) depicts the adjacent CpG loci (marked as 1 and 2), in which thymine originating from not methylated cytosine is present at cytosine within base CpG site, while cytosine that must have been methylated before bisulfite modification is present within consecutive CpG site.

Discussion

Recent ultra-deep NGS sequencing experiments again confirmed [3] a strong correlation of the methylation status of CpG sites spaced less than 50 bp [4], and it is generally accepted that the maintenance of co-methylation status of the adjacent CpG sites is essential for the regulatory function of CGIs. We analyzed the dynamics of the methylation changes at adjacent CpG sites across human genome in six different tissues and six blood cell types. The results of our analysis confirm, that methylation at the vast majority of adjacent CpG loci follows the principle of co-methylation. However, we also identified a subset of the adjacent CpG loci that display discordant and highly heterogeneous between specialized cell types methylation patterns. These CpG sites are not attribute of the bordering regions of CGIs or DMRs but are most significantly enriched in regions of the genome referred to as open sea.

Our results furthermore, suggest that discordantly methylated CpG sites display a high level of heterogeneity of methylation patterns between specialized cells are associated with genes providing specific cell functions. Contrary, to the CpG sites which in our analysis did not change methylation status between cells (predominantly constituting CGIs) that mapped to genes involved in general cells physiology. In the support of that observation, we also found that developmentally close cells to large extent share the methylation patterns at the adjacent CpG sites which is not surprising as the function of those cells is very similar.

The majority of loci with adjacent CpG sites displaying highest level of heterogeneity between specialized blood cells in our analyses, mapped to gene enhancers that harbor motifs which are bound by TFs with bZIP and ETS binding domains. TFs with those domains have been shown to be essential regulators of the hematopoiesis [42, 43] e.g., NFIL3 (bZIP TF family), which was shown to control regulatory function of T cells [44], or PU.1 (ETS TF family) which is well-known lineage-specific transcription factor, shown to be indispensable for generation of all known hematopoietic precursors with lymphoid developmental potential [45]. The binding of these and other TFs can potentially depend on methylation pattern within loci harboring adjacent CpG sites in specific cell type. And changes of methylation patterns at these loci could provide a mechanism for cell specific regulation function of enhancers harboring discordantly methylated loci. Especially that increasing number of research reports, show that enhancers display much greater than other

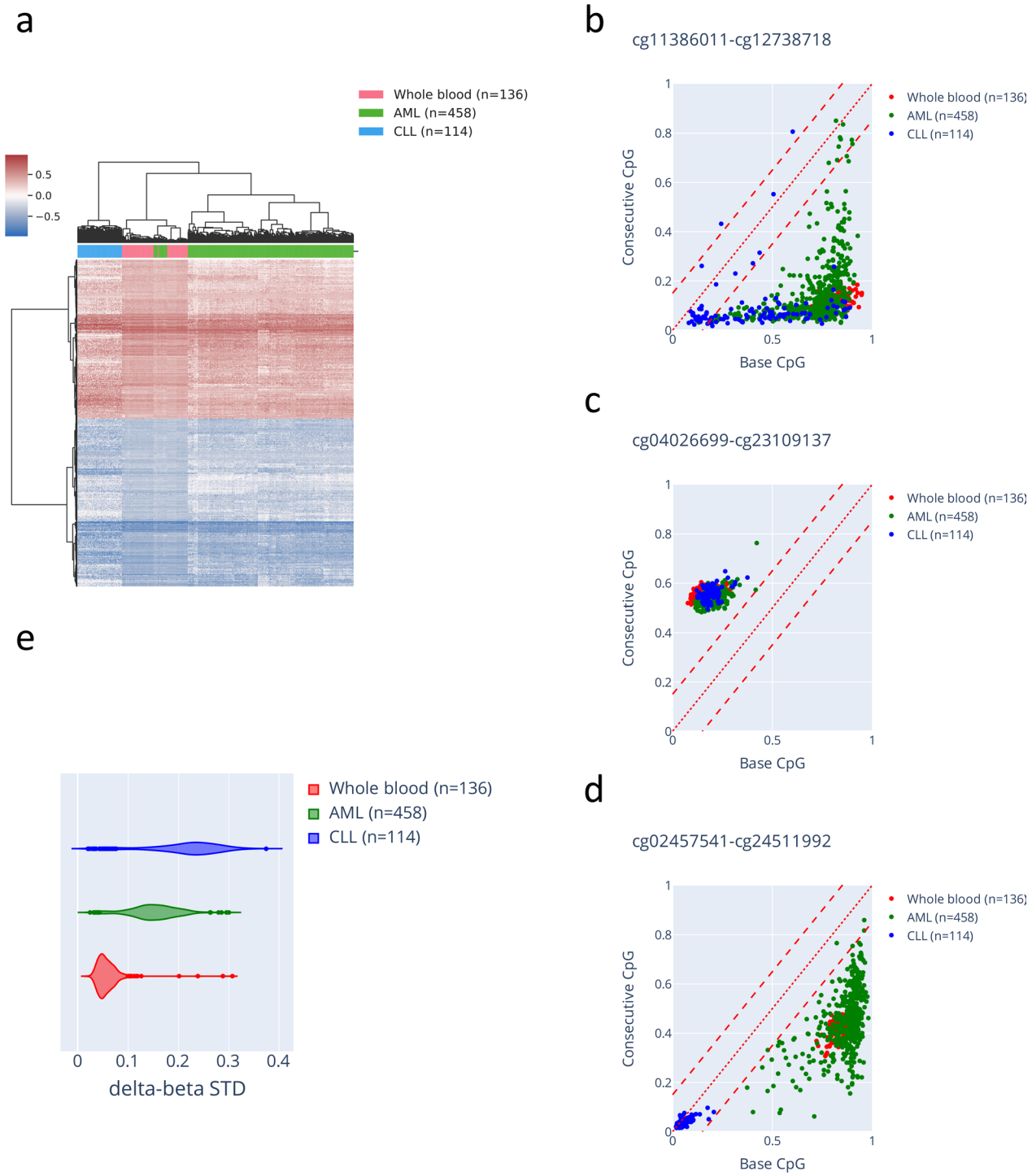

Fig. 7 Comparison of methylation patterns at adjacent CpG loci between healthy blood, acute myeloid leukemia (AML), and chronic lymphocytic leukemia (CLL). **a** Unsupervised hierarchical clustering of healthy blood, AML and CLL based on delta beta-values of CpG sites within "common discordantly methylated" loci (Supplementary Table 13). **b**–**d** Association of methylation levels at base and consecutive CpG site, with the solid line indicating no methylation level difference and dashed lines - ± 0.15 confidence interval, in **b** loci non co-methylated in whole blood, but displaying stochastic changes of the pattern in CLL and AML samples, **c** loci discordantly methylated in all sample types, and **d** loci discordantly methylated in whole blood, but co-methylated (at the level of 0%) in CLL, and stochastically changed in AML samples. **e** Comparison of standard deviation of delta beta-values of CpG sites within analyzed loci, between healthy blood AML, and CLL

regulatory regions differences in methylation level between specialized cell and tissue types, as well as developmental stages of cells [46–49]. Moreover, allele specific methylation of two pluripotency super-enhancers (i.e., Sox2 and the Mir290) was shown to switch in embryonic stem cells, resulting in cell-to-cell transcriptional heterogeneity [49].

The major limitation of our study is data resolution. Our results suggest that methylation patterns at the adjacent CpG sites reflect different methylation of the specific alleles, but similarly to previously identified cell type specific methylation changes not driven by the genotype [14]. However, only sequencing data for specific alleles and at single cell resolution would allow to in detail research the phenomenon we described. This type of data is still to be generated, but the new technologies, that allow sequencing of methylation patterns of long stretches of native DNA, with no need for bisulfite modification will provide this type of data. Also, mechanisms of interaction of DNA binding proteins with methylation patterns at the adjacent CpG loci during development, DNA replication and after cell division is a subject of future functional biology studies.

Due to the obvious technological limitations, we based our study on blood cells and similar studies need to be performed for other tissues and include all cells types that build a specific tissue. Nevertheless, similar results as we reported here for blood cell types, are expected in the analysis of different tissue types, but these analysis will likely identify different subset of discordantly methylated loci from those we found in specific blood cells.

Conclusions

In summary, EPIC microarray was designed on the principle of co-methylation and to target CpG sites spanning gene functional elements. Nevertheless, using this technology with all its limitations, we have been able to identify a substantial number of adjacent CpG loci that display remarkably stable blood cell type specific pattern of discordant methylation. The remarkable heterogeneity of the cell specific methylation patterns at those loci may provide a mechanism for altered binding of the DNA binding proteins and thus be involved in regulation of the activity of enhancers.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13072-024-00555-5.

Additional file 1

Additional file 2

Author contributions

O.T.L., J.B., K.S., and K.B. collected and produced primary material; O.T.L., J.B., and K.B., produced and analyzed sequencing data; O.T.L. and J.B. collected the data; O.T.L., J.B., P.P., K.E.S., and T.K.W analyzed the data; J.B. and O.T.L. prepared figures; O.T.L. and T.K.W. wrote the first draft of the paper and coordinated input of other authors to the text. All authors contributed to interpretation of findings and approved the final version of the manuscript.

Funding

This study was funded by OPUS grant from National Science Centre, grant ID: 2021/43/B/NZ2/02979 and Polish Returns grant from Polish National Agency for Academic Exchange, grant ID: PPN/PPO/2018/1/00088/.

Availability of data and materials

All data are available on the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) and Road Map Epigenomics Project (https://egg2.wustl.edu/roadmap/web_portal/index.html). Python and R packages used in this project are freely available from Bioconductor, CRAN, and Python Package Index. Homer tool is available from http://homer.ucsd.edu/homer/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Patent application has been filed for aspects of the technology used in the study.

Received: 30 May 2024 Accepted: 27 September 2024
Published online: 10 October 2024

References

1. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. 2000;9(16):2395–402.
2. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987;196(2):261.
3. Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, Miele G, Chiariotti L, Cocozza S. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics. 2020;112(1):144–50.
4. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378–85.
5. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet. 2017;49(4):635–42.
6. Haerter JO, Lövkvist C, Dodd IB, Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. Nucleic Acids Res. 2014;42(4):2235.
7. Hu K, Ting AH, Li J. BSPAT: a fast online tool for DNA methylation co-occurrence pattern analysis based on high-throughput bisulfite sequencing data. BMC Bioinform. 2015;16:1.
8. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44(11):5123.
9. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38(1):23–38.
10. Hannon E, Mansell G, Walker E, Nabais MF, Burrage J, Kepa A, Best-Lane J, Rose A, Heck S, Moffitt TE, et al. Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology. PLoS Genet. 2021;17(3):e1009443.

11. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315–22.
12. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480(7378):490–5.
13. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet. 2014;46(1):17–23.
14. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613(7943):355–64.
15. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, Stevenson K, Sougnez C, Wang L, Li S, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell. 2014;26(6):813.
16. Klughammer J, Kiesel B, Roetzer T, Fortelny N, Nemc A, Nenning KH, Furtner J, Sheffield NC, Datlinger P, Peter N, et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat Med. 2018;24(10):1611.
17. Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, Horn-Saban S, Zalcenstein DA, Goldfinger N, Zundelevich A, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet. 2012;44(1):1207.
18. Johnson KC, Anderson KJ, Courtois ET, Gujar AD, Barthel FP, Varn FS, Luo D, Seignon M, Yi E, Kim H, et al. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. Nat Genet. 2021;53(10):1456.
19. Gaiti F, Chaligne R, Gu H, Brand RM, Kothen-Hill S, Schulman RC, Grigorev K, Risso D, Kim KT, Pastore A, et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature. 2019;569(7757):576.
20. Zaimi I, Pei D, Koestler DC, Marsit CJ, De Vivo I, Tworoger SS, Shields AE, Kelsey KT, Michaud DS. Variation in DNA methylation of human blood over a 1-year period using the Illumina MethylationEPIC array. Epigenetics. 2018;13(10–11):1056–71.
21. Cubellis MV, Pignata L, Verma A, Sparago A, Del Prete R, Monticelli M, Calzari L, Antona V, Melis D, Tenconi R, et al. Loss-of-function maternal-effect mutations of PADI6 are associated with familial and sporadic Beckwith-Wiedemann syndrome with multi-locus imprinting disturbance. Clin Epigenetics. 2020;12(1):139.
22. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19(1):64.
23. Bolouri H, Farrar JE, Triche T Jr, Ries RE, Lim EL, Alonzo TA, Ma Y, Moore R, Mungall AJ, Marra MA, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med. 2018;24(1):103–12.
24. Hussmann D, Starnawska A, Kristensen L, Daugaard I, Thomsen A, Kjeldsen TE, Hansen CS, Bybjerg-Grauholm J, Johansen KD, Ludvigsen M, et al. IGHV-associated methylation signatures more accurately predict clinical outcomes of chronic lymphocytic leukemia patients than IGHV mutation load. Haematologica. 2022;107(4):877–86.
25. Park JL, Jeon S, Seo EH, Bae DH, Jeong YM, Kim Y, Bae JS, Kim SK, Jung CK, Kim YS. Comprehensive DNA methylation profiling identifies novel diagnostic biomarkers for thyroid cancer. Thyroid. 2020;30(2):192–203.
26. Oltra SS, Pena-Chilet M, Vidal-Tomas V, Flower K, Martinez MT, Alonso E, Burgues O, Lluch A, Flanagan JM, Ribas G. Methylation deregulation of miRNA promoters identifies miR124-2 as a survival biomarker in breast cancer in very young women. Sci Rep. 2018;8(1):14373.
27. Wang T, Maden SK, Luebeck GE, Li CI, Newcomb PA, Ulrich CM, Joo JE, Buchanan DD, Milne RL, Southey MC, et al. Dysfunctional epigenetic aging of the normal colon and colorectal cancer risk. Clin Epigenetics. 2020;12(1):5.
28. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189–96.
29. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, Teschendorff AE. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33(24):3982–4.
30. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4): e22.
31. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics (Oxford, England). 2016;32(4):587.
32. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.
33. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576.
34. Lahiri DK, Schnabel B. DNA isolation by a rapid method from human blood samples: effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality. Biochem Genet. 1993;31(7–8):321–8.
35. Calle-Fabregat C, Morante-Palacios O, Ballestar E. Understanding the relevance of DNA methylation changes in immune differentiation and disease. Genes (Basel). 2020;11(1):110.
36. Varrault A, Dubois E, Le Digarcher A, Bouschet T. Quantifying genomic imprinting at tissue and cell resolution in the brain. Epigenomes. 2020;4(3):21.
37. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, et al. The landscape of genomic imprinting across diverse adult human tissues. Genome Res. 2015;25(7):927–36.
38. Nan X, Meehan RR, Bird A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res. 1993;21(21):4886–92.
39. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
41. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
42. Tsukada J, Yoshida Y, Kominato Y, Auron PE. The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. Cytokine. 2011;54(1):6–19.
43. Ciau-Uitz A, Wang L, Patient R, Liu F. ETS transcription factors in hematopoietic stem cell development. Blood Cells Mol Dis. 2013;51(4):248–55.
44. Kim HS, Sohn H, Jang SW, Lee GR. The transcription factor NFIL3 controls regulatory T-cell function and stability. Exp Mol Med. 2019;51(7):1–15.
45. Rothenberg EV, Hosokawa H, Ungerback J. Mechanisms of action of hematopoietic transcription factor PU.1 in initiation of T-cell development. Front Immunol. 2019;10:228.
46. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, Ren B. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013;45(10):1198–206.
47. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.
48. Bogdanovic O, Smits AH, de la Calle ME, Tena JJ, Ford E, Williams R, Senanayake U, Schultz MD, Hontelez S, van Kruijsbergen I, et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet. 2016;48(4):417–26.
49. Song Y, van den Berg PR, Markoulaki S, Soldner F, Dall'Agnese A, Henninger JE, Drotar J, Rosenau N, Cohen MA, Young RA, et al. Dynamic enhancer DNA methylation as basis for transcriptional and cellular heterogeneity of ESCs. Mol Cell. 2019;75(5):905–20.

Publisher's Note