Supplemental Material to: "PoreMeth2 for decoding the evolution of methylome alterations with Nanopore sequencing."

Gianluca Mattei[1,*,+], Marta Baragli[1,*,+], Barbara Gega[1], Alessandra Mingrino[1], Martina Chieca[2], Tommaso Ducci[1], Gianmaria Frigè[3], Luca Mazzarella[3], Romina D'Aurizio[4], Francesco De Logu[2], Romina Nassini[2], Pier Giuseppe Pelicci[3,5], and Alberto Magi[1,+]

[1]Department of Information Engineering, University of Florence, Florence, Italy.

[2]Department of Health Science, Clinical Pharmacology and Oncology Section, University of Florence, Florence, Italy.

[3]Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milano, Italy.

[4]Institute of Informatics and Telematics (IIT), CNR, Pisa, Italy.

[5]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy.

[*]These authors contributed equally to this work.

[+] **Corresponding**: Alberto Magi, albertomagi@gmail.com. Gianluca Mattei, gianluca.mattei@unifi.it. Marta Baragli, marta.baragli@unifi.it

AML Samples Library Preparation and Sequencing

Diagnosis and relapse bone marrow samples were obtained from each AML patient. Mononuclear cells (MNCs) were isolated using Ficoll-Paque Plus density gradient centrifugation from samples exhibiting approximately 70% blast infiltration. Following separation, residual erythrocytes were removed by treatment with a lysis buffer composed of 155 mM NH_4Cl, 10 mM $KHCO_3$, and 0.1 mM Na_2EDTA. Genomic DNA was subsequently extracted from the purified MNC fraction.

DNA was qualitatively and quantitatively analyzed using Nanodrop and Qubit respectively. The sequencing was performed on the GridION 5X platform (Oxford Nanopore Technologies, Oxford, UK). Library preparation was performed according to the Oxford Nanopore Technologies (ONT) manufacturer's protocol for genomic DNA, the Ligation Sequencing Kit 1D SQK-LSK109. For each library 1 μg

of DNA as starting material was used. The input DNA for all nanopore libraries was unshared. Initially DNA was mixed with 3,5 ul of NEBNext FFPE DNA Repair Buffer and 2 μl of NEBNext FFPE DNA Repair Mix (M6630, NEB, Ipswich, Massachusetts) in order to repair possible nicks, and then end-prepped adding 3,5 ul of Ultra II End-prep reaction buffer and 3 μl of Ultra II End-prep enzyme mix (E7546, NEB). The reaction was incubated at 20°C for 5 minutes and 65°C for 5 minutes, followed by a purification step with 60 μl of 1X AMPure XP beads, by a washing and an elution. The end-repaired DNA was ligated with 5 μl Adapter Mix (AMX, ONT) using 10 μl NEBNext Quick T4 DNA Ligase (E6056, NEB) for 10 min at room temperature. The adapted-ligated DNA was cleaned up by adding 40 μl of 1X AMPure XP beads and incubated at room temperature for 5 min. The beads were pelleted on a magnetic rack and pellet was washed twice by resuspending in 250 μl Long Fragment Buffer (LFB, ONT) in order to enrich for long molecules. Then, the reaction was eluted in 15 μl of Eluition Buffer (EB, ONT) at room temperature for 10 minutes. The prepared library was quantified by Qubit and then, prior to loading onto the flow cell, was mixed with 37,5 μl of Sequencing Buffer (SB, ONT) and 25,5 μl of Loading Beads (LB, ONT). According to the ONT's protocol before loading, each R9.4.1 SpotON flow cell (FLO-MIN106D) was primed mixing the Flush Tether (FLT, ONT) directly to the tube of Flush Buffer (FB, ONT) and loading this mix via the priming port. Libraries were sequenced on a GridION 5x device and, in order to maximize the total throughput, sequencing was carried out for 48 h.

HPGC Samples Library Preparation and Sequencing

Human peripheral glial cells (HPGC) (#P10351, Innoprot) were cultured and maintained in dedicated medium (P60123, Innoprot) at 37°C in 5% CO2 and 95% O2. HPGC were cultured in a 6 cm Petri dish until the 80% of confluence and the day after were treated for 0 (T0) or 48 (t48) hours with G protein-coupled receptor agonist (S1198, Merck). Genomic DNA and total RNA were then extracted using DNeasy Blood & Tissue Kit (69506, Qiagen) and RNeasy Mini Kit (74106, Qiagen) respectively. Before Nanopore Library preparation, total RNA was quantified via Qubit Fluorometer with RNA HS Assay Kit (32852, Thermo Fisher Scientific) and RNA quality was validated with Agilent 2100 Bioanalyzer using RNA 6000 Nano Kit (5067-1511, Agilent). For RNA and DNA library preparation, the PCR-cDNA Barcoding Kit (SQK-PCB111.24) and SQK-LSK109-XL kit respectively were used following manufacturer's instructions.

Bivariate SLM

For DMRs identification we developed a bivariate version of the SLM algorithm.

$\Delta\beta$ and ΔS signals for N CpGs are modeled as a sequential processes $x = (x_1, .., x_i, .., x_N)$ (with

x_i=($\Delta\beta_i, \Delta S_i$)) as the sum of two independent stochastic processes:

$$x_i = m_i + \epsilon_i \tag{1}$$

$$m_i = (1 - z_i) \cdot m_{i-1} + z_{i-1} \cdot (\mu + \delta_i) \tag{2}$$

where :

- $m_i = (m_{i1}, m_{i2})$ is the vector of the unobserved mean level,
- ϵ_i is the vector of white noises and it follows a bivariate normal distribution with mean $\mu\epsilon = [0, 0]$ and covariance matrix $\sum_\epsilon$ ($\epsilon_i \sim N(0, \Sigma_\epsilon)$),
- z_i are random variables taking the values in [0,1] with probabilities (η = Pr(z_i = 1), $(1 - \eta) = Pr(z_i = 0)$,
- δ_i are random vectors that follow a bivariate normal distribution with mean μ and covariance matrix Σ_μ.

Moreover, m_i depends by z_i: when $z_{i-1} = 0$, m_i is the same as m_{i-1} and when $z_{i-1} = 1$, m_i takes its new value according to a bivariate Gaussian law with mean μ and covariance matrix $\sum_\epsilon$ independently of m_{i-1} :

As x_i are modeled as the sum of two independent stochastic processes, the expected value of x_i is μ, and its covariance matrix is the sum of the covariances of the two processes:

$$E(x_i) = \mu = (\mu_{\Delta\beta}, \mu_{\Delta S}) \tag{3}$$

$$Cov(x_i) = \Sigma_\mu + \Sigma_\epsilon \tag{4}$$

Hence, we can introduce a different parametrization of the SLM by defining the parameter ω such that $\Sigma_\mu = \omega \cdot \Sigma$ and $\Sigma_\epsilon = (1–\omega) \cdot \Sigma$.

We also hypothesize that the white noise distributions and mean level distributions can be considered independent between $\Delta\beta$ and ΔS signals:

$$N(0, \Sigma_\epsilon) = N(0, \sigma_{\epsilon_{\Delta\beta}}) \cdot N(0, \sigma_{\epsilon_{\Delta S}}) \tag{5}$$

$$N(\mu, \Sigma_\mu) = N(\mu_{\Delta\beta}, \sigma_{\mu_{\Delta\beta}}) \cdot N(\mu_{\Delta S}, \sigma_{\mu_{\Delta S}}) \tag{6}$$

where $\sigma_{\epsilon_{\Delta\beta}}$ and $\sigma_{\epsilon_{\Delta S}}$ are the standard deviations of the white noises of $\Delta\beta$ and ΔS signals, while $\sigma_{\mu_{\Delta\beta}}$ and $\sigma_{\mu_{\Delta S}}$ are the standard deviations of the mean level distributions.

With these assumptions, the random process z_i is the only variable correlating the two signals. When z_i changes its value, the mean level of $\Delta\beta$ and ΔS shifts, allowing the joint model to detect common shifts in the mean Magi et al. 2011.

As a consequence of the previous definitions, the joint distribution of the observations and latent variables, given the parameters, has this form:

$$\begin{aligned} p(x, m, z|\theta) &= p(x|m, \Sigma_\epsilon) \cdot p(m|z, \mu, \Sigma_\mu) \cdot p(z|\eta) \\ &= \prod_{i=1}^{N} p(x_i|m_i, \Sigma_\epsilon) \cdot p(m_0) \cdot \prod_{i=0}^{N-1} p(m_{i+1}|m_i, z_i, \mu, \Sigma_\mu) \cdot p(z_i|\eta) \end{aligned} \tag{7}$$

where $\Theta = (\mu, \Sigma_\mu, \Sigma_\epsilon, \eta)$ and:

- $p(x_i|m_i, \Sigma_\epsilon) = \mathcal{N}(x_i|m_i, \Sigma_\epsilon)$ is the probability distribution of x_i given m_i and the parameters.
- $p(m_{i+1}|m_i, z_i, \mu, \Sigma_\mu) = (1 - z_i) \cdot \delta(m_{i+1} - m_i) + z_i \cdot \mathcal{N}(m_{i+1}|\mu, \Sigma_\mu)$ is the probability distribution of the latent variable m_{i+1} given m_i, z_i, and the parameters; δ is the Dirac delta function.
- $p(z_i|\eta) = \eta \cdot \delta(z_i - 1) + (1 - \eta) \cdot \delta(z_i)$ is the probability density function of z_i.

Equation 7 defines a Hidden Markov Model (HMM) of order one, in which a single state variable, $q_i = (m_i, z_i)$, summarizes all the relevant past information of the underlying process. In the model defined by SM1 the elements of the HMM are the following:

- The state transition probability distribution is: $p(q_{i+1}|q_i, \theta) = p(m_{i+1}|m_i, z_i, \mu, \Sigma_\mu) \cdot p(z_i|\eta)$
- The emission probability distribution is: $p(x_i|q_i, \theta) = p(x_i|m_i, \Sigma_\epsilon)$
- The initial state probability distribution is: $p(q_0|\theta) = p(m_0|\mu, \Sigma_\mu)$

Bivariate SLM algorithm

The joint distribution of Equation 7 defines an Hidden Markov Model (HMM) of order one with state variable $q_i = (m_i, z_i)$ and bivariate emission probability. The fact that the multivariate SLM is an HMM allows us to make use of the several algorithms developed for these kinds of models.

To maximize the likelihood of the multivariate extension of shifting level model we use a procedure similar to that used in (Magi et al. 2017). We introduce a markovian stochastic process $s_1, ..., s_k$ taking

values in $S = 1, 2, ..., K$. We assume that the conditional probability of x_i, given $s_i = k$, is a bivariate truncated normal with mean $\mu_k = (\mu_{1k}, \mu_{2k})$ and variance Σ_ϵ, and the parameter μ_k is associated to each state of the markovian stochastic process and represents an approximation of the m_i latent variables of the bivariate SLM.

Remembering equations 7 and 6 and that $\Delta\beta$ and ΔS take values in the range [-1,1], we assume that the conditional probability of x_i, given $s_i = k$ is a truncated gaussian distributions with upper and lower bound 1 and -1 respectively, the emission probability distribution has the following form:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_\epsilon}\right)^2\right] \cdot \frac{1}{\Phi(\frac{1-\mu_k}{\sigma_\epsilon}) - \Phi(\frac{-1-\mu_k}{\sigma_\epsilon})}. \tag{8}$$

where Φ represent the cumulative distribution functions of a non-truncated Gaussian of parameter μ_k and σ_ϵ.

To complete the description of the model, it remains to specify the state transition matrix P. From (7) the state transition probability has the following form:

$$\begin{aligned} p(m_{i+1}|m_i, z_i, \mu, \sigma^2, \omega) \cdot p(z_i|\eta) \quad &= \quad [(1-z_i)\cdot\delta(m_{i+1}-m_i) + z_i \cdot N(m_{i+1}|\mu, \omega\cdot\sigma^2)]\cdot \\ &\quad \cdot[\eta\cdot\delta(z_i-1) + (1-\eta)\cdot\delta(z_i)], \end{aligned} \tag{9}$$

hence the state transition matrix is:

$$P_{jk} = \begin{cases} (1-\eta) + \eta \cdot g_{jk} & j = k \\ \eta \cdot g_{jk} & j \neq k \end{cases} \tag{10}$$

where

$$\begin{aligned} g_{jk} &= c_j \cdot e^{-\frac{(\mu_k-\mu)^2}{2\sigma_\mu^2}}, \\ c_j &= \left(\sum_{k=1}^{K} e^{-\frac{(\mu_k-\mu)^2}{2\sigma_\mu^2}}\right)^{-1}. \end{aligned} \tag{11}$$

To estimate the parameters of the truncated-gaussian bivariate Shifting Level Model (BiSLM), we develop a two-step algorithm that follows our previous idea in (Magi et al. 2013). Since the$\Delta\beta$ and ΔS can take values in a well-defined range, we used a large number of states K_0 and we choose $\mu_k = (\mu_{k2}, \mu_{k2})$ in order to densely and homogeneously cover al the combination of the ranges ([-1,1],[-1,1]) instead of estimating the μ_k parameters by using the Baum and Welch algorithm. Moreover, since we expect that the great majority of CpG have neither differential methylation nor entropy between samples ($\Delta\beta \sim 0$,$\Delta S \sim 0$), we initialize $\mu = (0, 0)$. This simple solution drastically improves the computational

performance of our algorithm without affecting its accuracy in the detection of signal shifts.
In the first step of the algorithm we initialize the mean μ and the variances σ^2, σ_μ^2 and σ_ϵ^2 with the following formulas:

$$\begin{aligned} &\mu = (0,0), \\ &\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{(N-1)}}, \\ &\sigma_\mu^2 = \omega \cdot \sigma^2, \\ &\sigma_\epsilon^2 = (1-\omega) \cdot \sigma^2. \end{aligned} \tag{12}$$

In the second step we apply the Viterbi algorithm to find the best state sequence $s^{(j)}$ and estimate the points of mean shift z_i. After Viterbi algorithm we calculate the median of the $\Delta\beta$ and ΔS values that belong to each segment.

The inputs to the algorithm are the $\Delta\beta$ and ΔS values ($\Delta\beta = (\Delta\beta_1, .., \Delta\beta_i, .., \Delta\beta_N)$, $\Delta S = (\Delta S_1, .., \Delta S_i, .., \Delta S_N)$) to be segmented, number of states K_0, the parameter ω and the parameter θ.

Synthetic Data Generation

Synthetic methylation profiles were simulated by first generating long methylation patterns, made of N CpG dinucleotides, as a sequence of consecutive short epialleles (made of three, see Supplemental Figure S1) randomly sampled from multinomial probability distributions that mimic methylated and unmethylated epialleles. The multinomial distribution models the probability of counts of the eight possible three-CpG epialleles ('111','110','101',") with the parameter $p = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$. By using multinomial probability distributions with different values of p it is possible to sample methylated or unmethylated epialleles with different level of randomness in different position of the pattern (see Supplemental Figure S1).

To mimic real methylation patterns, mainly made of methylated CpGs, consecutive short epialleles were sampled from multinomial distribution with $p = (p_1 = 0.9, p_2 = (1-p_1)/3, p_3 = (1-p_1)/3, p_4 = (1-p_1)/3, p_5 = 0, p_6 = 0, p_7 = 0, p_8 = 0)$. We then added, in specific position of the pattern, consecutive epialleles sampled from multinomial distribution with different values of p.

Low-entropy methylated epialleles were generated using $p_5 = p_6 = p_7 = p_8 = 0$, $p_1 > 0.5$ and $p_2 = p_3 = p_4 = (1-p_1)/3$. Similarly, low-entropy unmethylated epialleles were generated using $p_1 = p_2 = p_3 = p_4 = 0$, $p_5 > 0.5$ and $p_6 = p_7 = p_8 = (1-p_5)/3$. High-entropy methylated epialleles are generated with $p_5 = p_6 = p_7 = p_8 = 0$ and $p_1 = p_2 = p_3 = p_4 = 0.25$, while high-entropy unmethylated

epialleles are generated with $p_5 = p_6 = p_7 = p_8 = 0.25$ and $p_1 = p_2 = p_3 = p_4 = 0$ (see Supplemental Figure S1.a).

By following these strategy we generated M methylation patterns for test and control samples that allow to simulate methylated or unmethylated epialleles with a predefined level of randomness (entropy). In particular we generated consecutive epialleles made of 5-50 CpGs (5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 CpGs) that simulate hyper- or hypo-methylation (between test and control) with hyper-, hypo- or iso-entropic changes (see Supplemental Figure S1.b).

The M methylation patterns were finally used to generate long reads by sampling segments of consecutive CpGs from the simulated patterns in order to obtain the desired sequencing coverage (see Supplemental Figure S1.c). To test the performance of BiSLM we simulated methylation patterns with epiallelic changes of different sizes (from five to 50 CpGs) and we sampled segments to generate sequencing coverages from 10x to 50x.

Synthetic Data Analysis with BiSLM, Modkit and BSmooth

In order to evaluate the ability of our BiSLM to identify DMRs of different size and with different epiallelic changes we simulated synthetic methylation profiles. To this end, we developed a computational recipe that simulates long methylation patterns, made of N CpG dinucleotides, by generating consecutive short epialleles randomly sampled from a multinomial probability distribution (see Methods and Supplemental Figure S2). Methylation patterns are then exploited to generate long reads datasets with desired sequence size and sequencing coverage. By using this strategy we simulated test and control samples at sequencing coverages that range from 10x to 50x (10, 20, 30, 40 and 50x) with hypo- and hyper-methylated DMRs from 5 to 50 CpGs (5, 10, 15, 20, 25, 30, 35, 40, 45, 50). hypo- and hyper-methylation was simulated with random and non-random epiallelic shifts and different level of randomness (different multinomial probability distribution, see caption of Supplemental Figure S2). For each pair of test and control samples we calculated frequency β and entropy S at each CpG and we then applied BiSLM to simultaneously segment $\Delta\beta$ and ΔS. A simulated DMR is considered a true positive if it has a reciprocal overlap larger than 0.9 with a region detected by BiSLM. As a first step we exploited synthetic data to evaluate the performance of our approach as a function of parameter settings (η and ω). Supplemental Figures S3-7 show that the best F-measure (trade off between precision and recall) is obtained for $\omega = 0.1 - 0.2$ and $\eta = 10^{-5} - 10^{-6}$ for all sequencing coverages (10x-50x). To further evaluate BiSLM's ability to detect DMRs of varying sizes and characterize their epiallelic changes, we analyzed synthetic data and compared its performance with Modkit and BSmooth. While BiSLM's was used to contemporary segment $\Delta\beta$ and ΔS signals, Modkit and BSmooth were applied to ΔS profiles. First, we computed the F1-score (a balance between precision and recall) for each method at different reciprocal overlap thresholds (0.1, 0.5,

and 0.9). As shown in Supplemental Figure S8, panels a–c, BiSLM maintained consistent performance across all overlap values, whereas the other two methods exhibited decreasing F1-scores as the threshold increased. These findings highlight BiSLM's superior precision in defining DMR boundaries. Next, we assessed F1-score performance as a function of DMR size (measured by the number of CpGs), using a reciprocal overlap threshold of 0.7. Again, BiSLM outperformed Modkit and BSmooth across all simulated sequencing coverages (Supplemental Figure S8, panel d), demonstrating its enhanced resolution capabilities. Overall, our synthetic data analysis underscores the impact of sequencing coverage on the accuracy of DMR identification. Detecting DMRs as small as five consecutive CpGs requires a minimum sequencing coverage of ≥ 20, while accurate classification of epiallelic formation demands even higher coverage (≥ 30). Notably, iso-entropic DMRs are more challenging to detect than hyper- and hypo-entropic alterations (Supplemental Figure S8, panel e).

Supplemental methods for AML dataset

As a first step, to investigate whether repetitive elements represent hard-to-map regions with potentially lower coverage, which could limit DMR detection by BiSLM algorithm, we compared the fraction of CpG sites as a function of sequencing coverage genome-wide and specifically within repetitive elements. The results reported in Supplemental Figure S9, indicate no significant differences in coverage between repetitive regions and the rest of the genome. These findings suggest that our DMR identification method performs consistently across both repetitive and non-repetitive regions. As a further step, in order to evaluate the performance of PoreMeth2 in the analyses of real cancer data we applied it to three AML sample pairs that we previously analyzed in (Magi et al. 2023), and we compared the DMRs identified by our method with those detected by Modkit and BSmooth. As a first step, to compare the results obtained from the three methods, we analyzed the number of DMRs and the distribution of DMR sizes (in terms of both base pairs and the number of CpGs), CpG density, and $\Delta\beta$ values (Supplemental Figures S13, S14 and S15). Our algorithm identified around three thousands DMRs for each sample pairs (3102 for UD5, 2825 for UD10 and 2874 for AML2). On the other hand, BSmooth detected just over a hundred DMRs per pair (116 for UD5, 166 for UD10, and 137 for AML2), whereas Modkit identified a large number of DMRs (4114 for UD5, 801 for UD10, and 1279 for AML2). The average size of DMRs detected by our tool ranged from 700 to 800bp across all three AML pairs, BSmooth identified DMRs with an average size of 300–500 bp, while Modkit exhibited a bimodal size distribution, with the first peak at just a few bases and the second peak at 100–200 bp (panels a of Supplemental Figures S13, S14 and S15). These results are also reflected in the number of CpGs per DMR. While PoreMeth2 and BSmooth identify DMRs composed of approximately 10 to 100 CpGs, the vast majority of DMRs identified by Modikit consist of fewer than 3–4 CpGs, with a peak at a single CpG (panels b of Supplemental Figures S13, S14

and S15). The great majority (60–70%) of the DMRs identified by our method and BSmooth exhibited a CpG density of ≤ 3 CpGs per 100 bp, whereas only $\sim 30-40\%$ of the DMRs detected by Modkit fell within this range (panels d of Supplemental Figures S13, S14 and S15). These results suggest that while BiSLM and BSmooth are capable of analyzing DMRs in both high- and low-CpG-density regions, Modkit has reduced detection capabilities in low-CpG-density regions, which constitute the vast majority of the epigenome. Moreover, our approach identified DMRs across a broad range of $\Delta\beta$ values, whereas the other two methods tended to detect DMRs with significantly more pronounced changes in methylation frequency. Specifically, BSmooth identified hypermethylated DMRs with an average $\Delta\beta = 0.5$ and hypomethylated DMRs at $\Delta\beta = -0.5$, while Modkit detected DMRs with even more extreme values, averaging $\Delta\beta = 0.8$ and $\Delta\beta = -0.8$. Surprisingly, Modkit also identified regions with $\Delta\beta$ values close to zero as DMRs (panels c of Supplemental Figures S13, S14 and S15). To assess the concordance between the three approaches, we analyzed the reciprocal overlap of the DMRs identified by each method in each of the three AML pairs. The overlap between BiSLM and BSmooth is nearly complete for all the three AML pairs, whereas the overlap with Modkit is considerably lower (Supplemental Figure S17.a-c). To evaluate the reliability of the DMRs uniquely identified by either BiSLM or Modkit, we analyzed them separately. Most DMRs detected exclusively by Modkit contain fewer than four CpGs ($\sim 63\%, 14\%, 5\%$ with 1, 2, 3 CpGs respectively in UD5, $\sim 34\%, 14\%, 13\%$ in UD10 and $\sim 48\%, 15\%, 8\%$ in AML2), and those with more than five CpGs display $\Delta\beta$ values close to zero (Supplemental Figure S17.d-f), suggesting the absence of true differential methylation. To assess the reliability of the DMRs uniquely identified by BiSLM, we recalculated the $\Delta\beta$ values and the likelihood ratio scores for each DMR using Modkit's DMR scoring method (see `https://nanoporetech.github.io/modkit/dmr_scoring_details.html#likelihood-ratio-scoring-details`). The results, shown in Supplemental Figure S18, demonstrate a very high correlation between the $\Delta\beta$ values calculated by Modkit and PoreMeth2 (Pearson's r = 0.99 across all three AML sample pairs). Moreover, approximately 99% of the DMRs exhibited a Modkit significance score greater than 5, suggesting that the DMRs identified by BiSLM are highly reliable. These findings highlight the superior capability of our method in accurately detecting DMRs across a broad spectrum of CpG densities and $\Delta\beta$ values. Moreover, they demonstrate that long-read sequencing, combined with our novel computational approach, enables the identification of epiallelic changes between test and control samples at an unprecedented resolution. As a final step, we compared the DMRs identified by BiSLM across the three AML sample pairs and observed limited overlap. Only 55 DMRs were common to all three samples, while 296 were shared between UD10 and AML2, 134 between UD5 and AML2, and 226 between UD10 and UD5 (Supplemental Figure S30). The 55 shared DMRs encompassed 65 genes, among which only three were DEGs in both UD5 and UD10, but not in AML2 (Supplemental Figure S31). Notably, two of these three DM-DEGs exhibited opposite Log2FC directions in the UD5

and UD10 samples (Supplemental Figure S32). These findings are consistent with our previous study, where we demonstrated that a limited number of DMRs, affecting sparse CpGs within gene bodies of a few, patient-specific master transcription factors, represent the primary drivers of gene deregulation associated with the chemo-resistant phenotype.

Supplemental methods for Human Peripheral Glial Cells dataset

As for AML dataset, to investigate whether repetitive elements represent hard-to-map regions with potentially lower coverage, which could limit DMR detection by BiSLM algorithm, we compared the fraction of CpG sites as a function of sequencing coverage genome-wide and specifically within repetitive elements. The results reported in Supplemental Figure S33, indicate no significant differences in coverage between repetitive regions and the rest of the genome. These findings suggest that our DMR identification method performs consistently across both repetitive and non-repetitive regions.

References

Magi A, Benelli M, Yoon S, Roviello F, and Torricelli F. 2011. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.* **39**: DOI: 10.1093/nar/gkr068.

Magi A, Pippucci T, and Sidore C. 2017. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics.* **18**: 747.

Magi A et al. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* **14**:

Magi A et al. 2023. High-resolution Nanopore methylome-maps reveal random hyper-methylation at CpG-poor regions as driver of chemoresistance in leukemias. *Commun Biol.* **6**: 382.

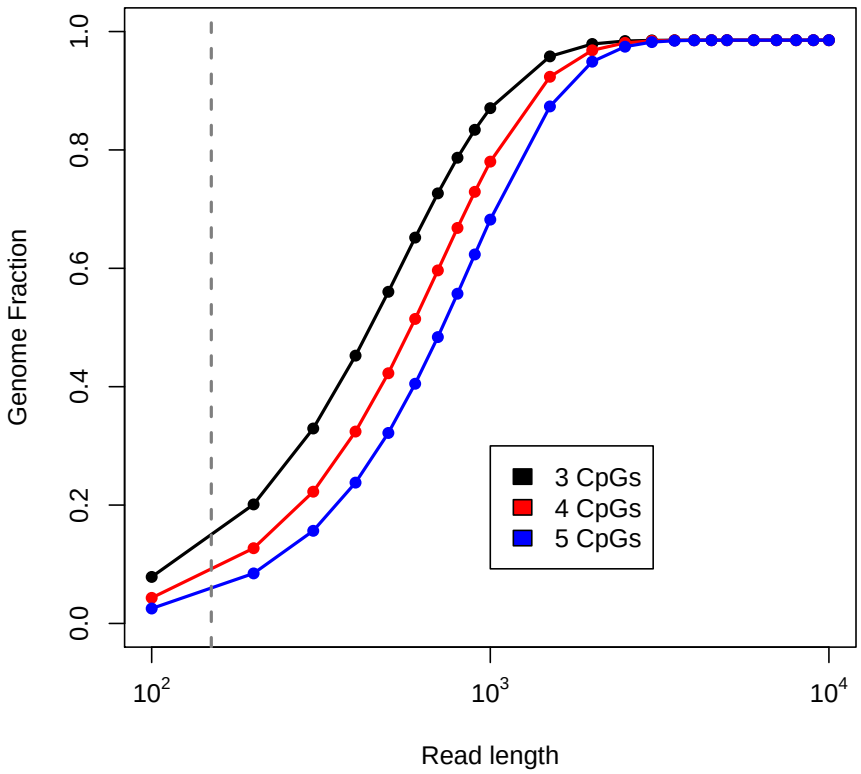

Figure S1: The plot shows the fraction of epialleles covered across the whole genome as a function of read length, spanning from 100 bp to 10 kb. The dotted line represents the mean read size by Illumina sequencing technology, highlighting how Second Generation Sequencing data only allow epialleles characterization in a small portion of the genome (10-15%). The blue, red and black lines display the distribution of the covered genome fraction when applying the bivariate SLM to simulated epialleles with different number of CpGs.

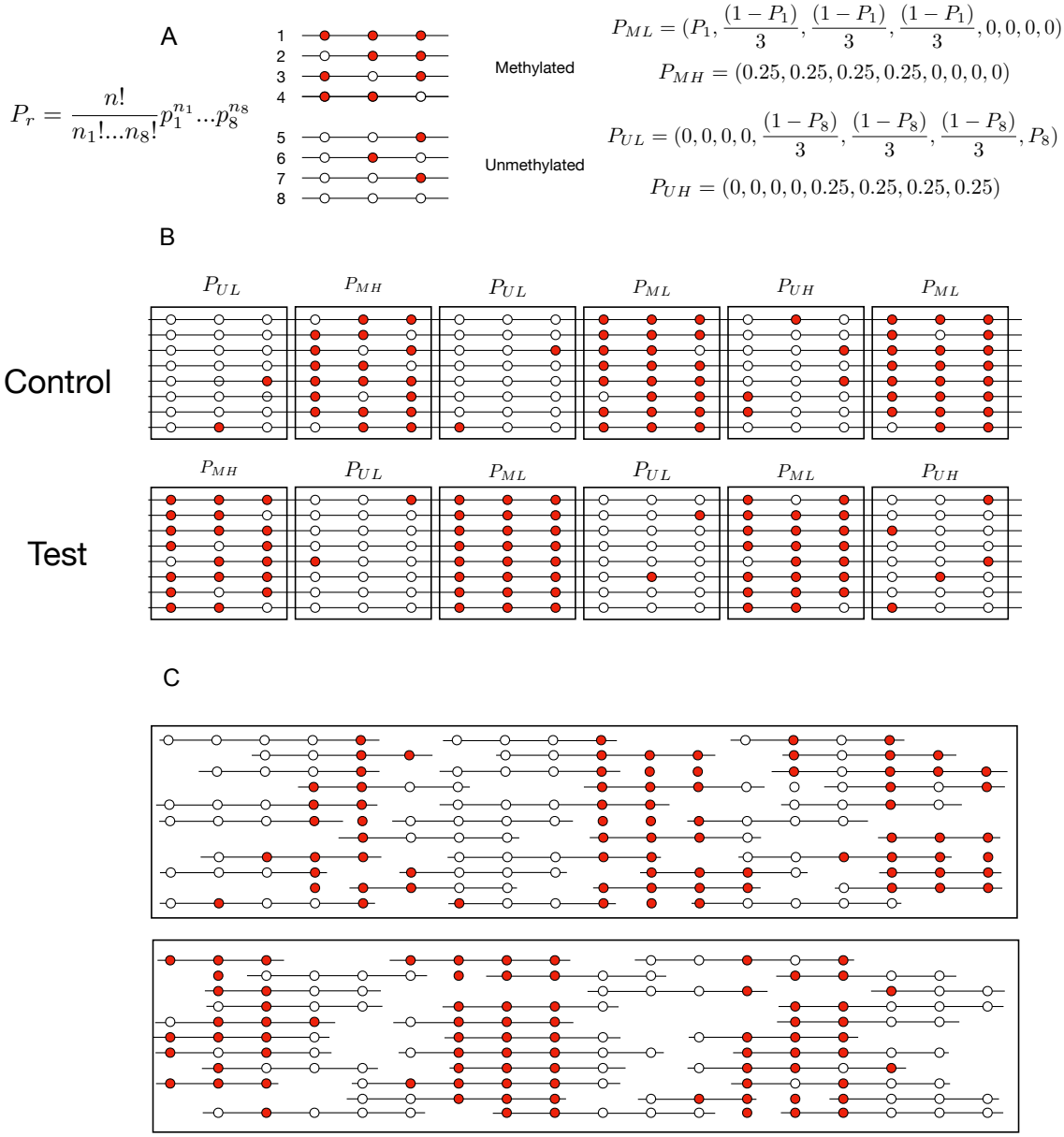

Figure S2: Computational scheme to generate synthetic methylation profiles. Methylation patterns are generated as a sequence of consecutive three-CpGs epialleles ('111','011','101','110') randomly sampled from multinomial probability distribution. To simulate methylated and unmethylated patterns with different level of randomness we used four distributions with parameter P_{ML} (methylated with low stochasticity), P_{MH} (methylated with high stochasticity), P_{UL} (unmethylated with low stochasticity) and P_{UH} (unmethylated with high stochasticity). Setting different values of P_1 and P_8 allows to modulate the stochasticity of the epialleles and consequently methylation entropy. In our simulations we set $P_1 > 0.5$ and P_8 with values ranging between 0.5 and 0.9. Following these scheme we generated methylation patterns for test and control samples that simulate all the six possible epiallelic changes (B). The final step consists in sampling segments of consecutive CpGs from the simulated patterns in order to obtain the desired sequencing coverage. To test the performance of BiSLM we simulated methylation patterns with epiallelic changes of different sizes (from five to 50 CpGs) and we sampled segments to generate sequencing coverages from 10x to 50x.

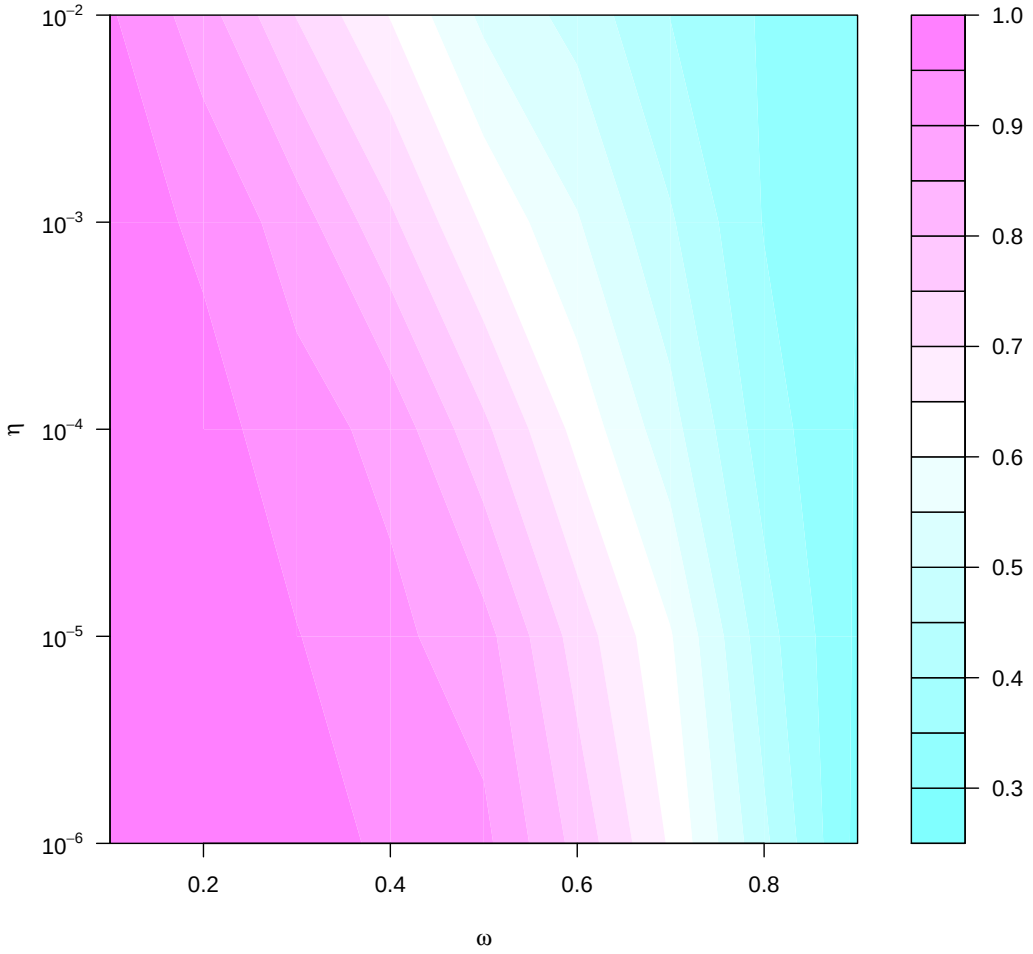

Figure S3: The contour plot shows the performance (in terms of F-Measure) of BiSLM for different combinations of values of η and ω parameters in the analysis of methylation profiles generated from synthetic epialleles at 10x of sequencing coverage.

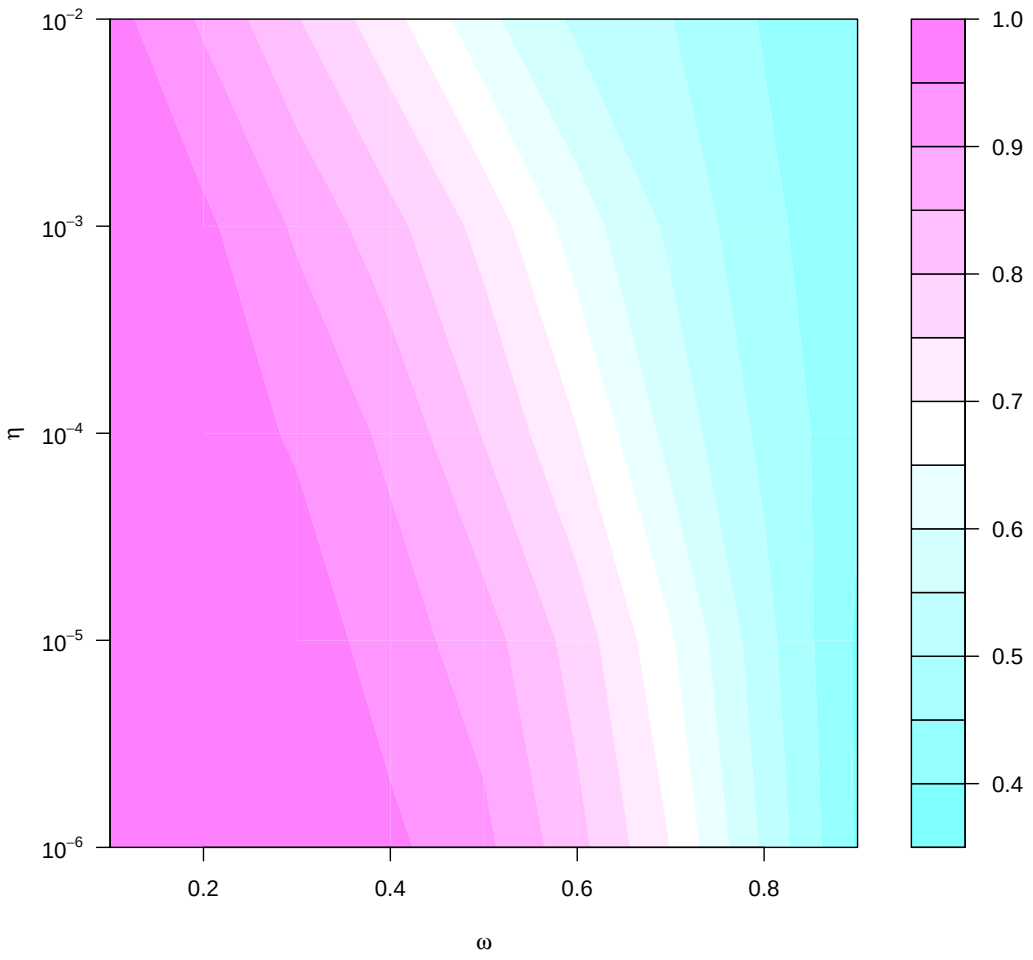

Figure S4: The contour plot shows the performance (in terms of F-Measure) of BiSLM for different combinations of values of η and ω parameters in the analysis of methylation profiles generated from synthetic epialleles at 20x of sequencing coverage.

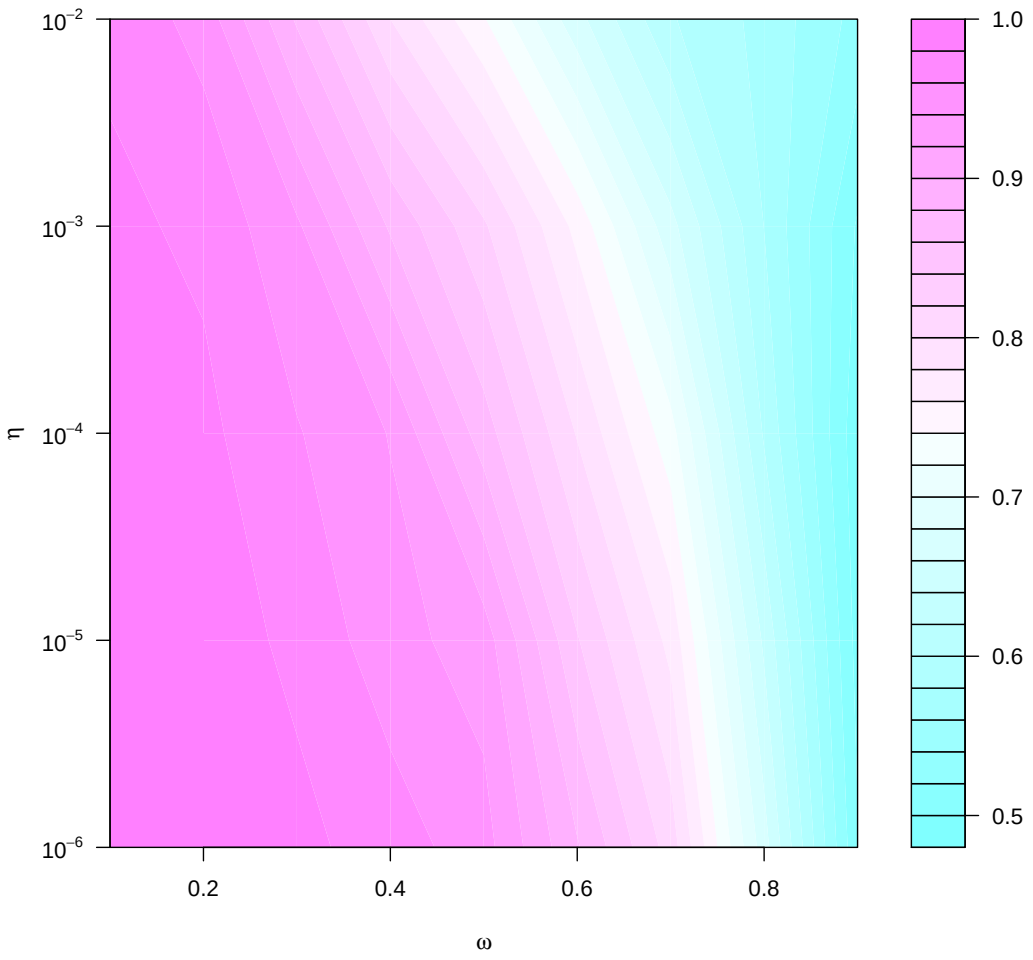

Figure S5: The contour plot shows the performance (in terms of F-Measure) of BiSLM for different combinations of values of η and ω parameters in the analysis of methylation profiles generated from synthetic epialleles at 30x of sequencing coverage.

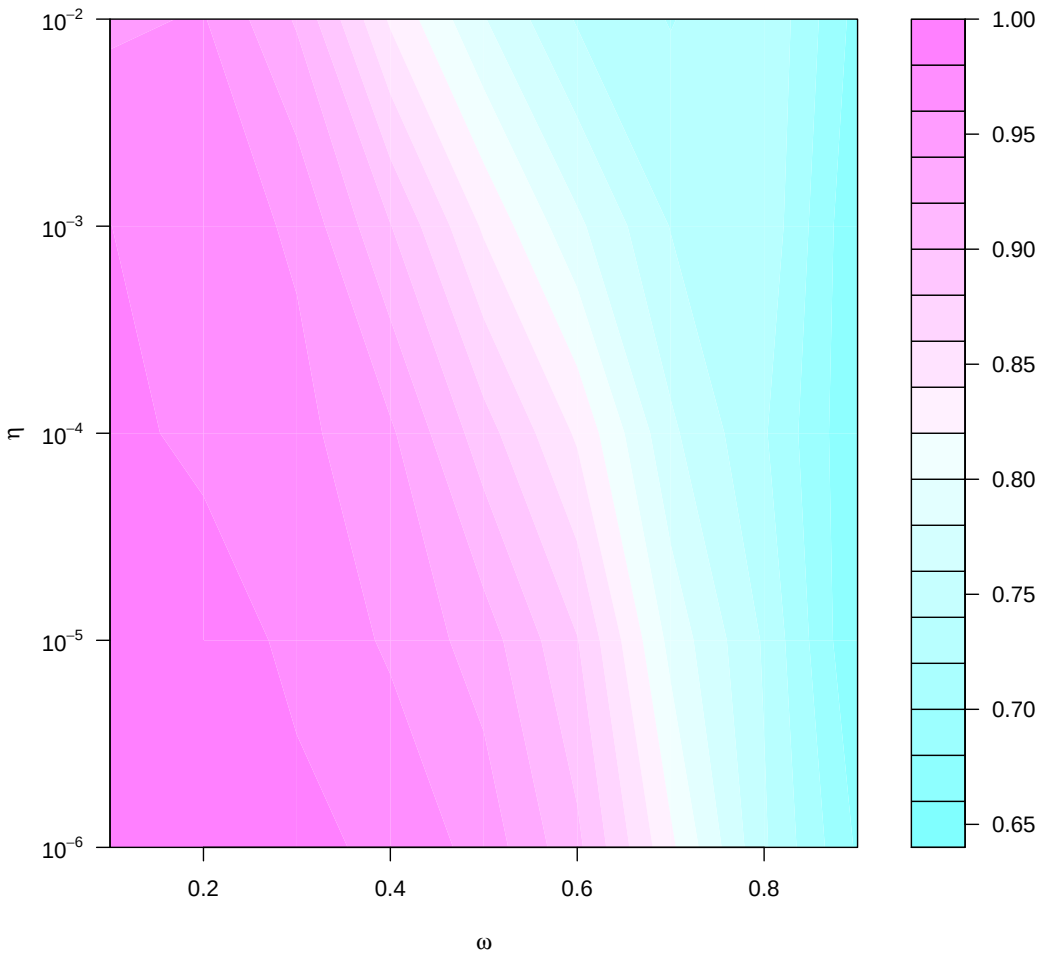

Figure S6: The contour plot shows the performance (in terms of F-Measure) of BiSLM for different combinations of values of η and ω parameters in the analysis of methylation profiles generated from synthetic epialleles at 40x of sequencing coverage.

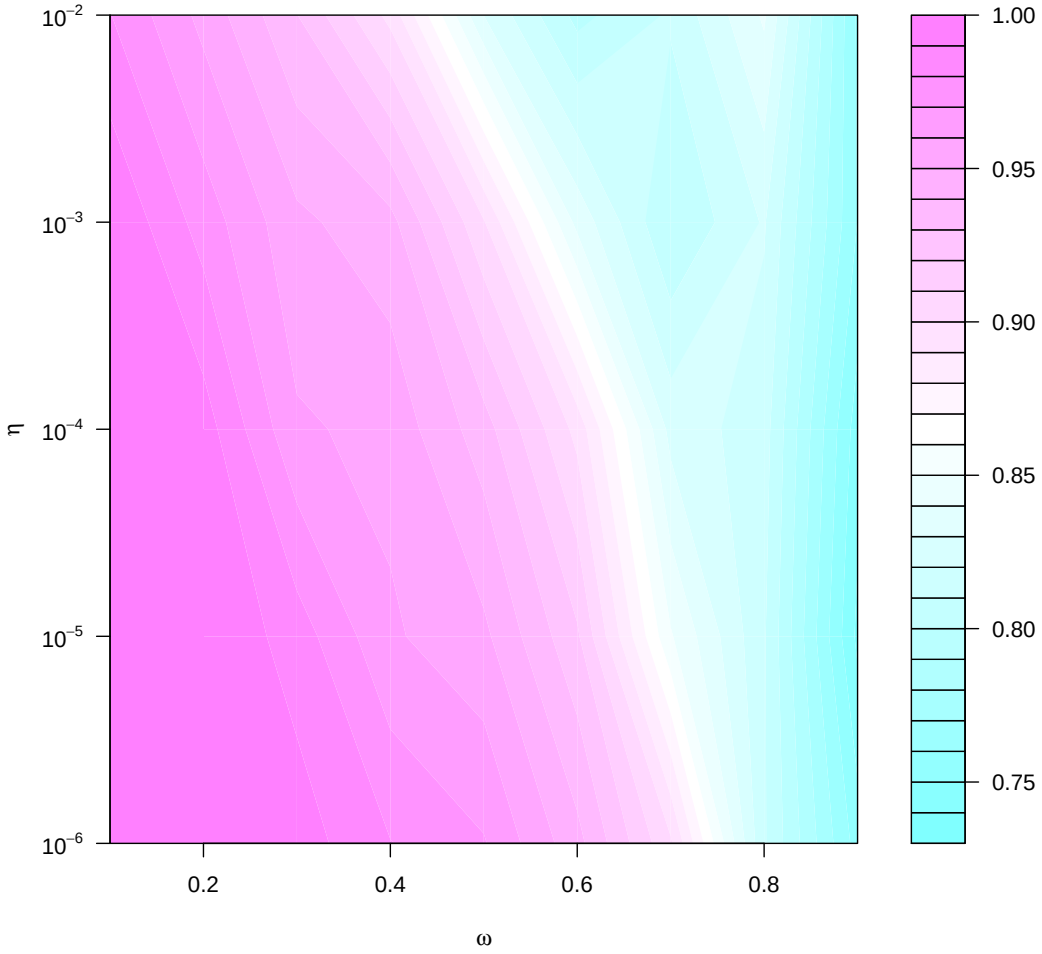

Figure S7: The contour plot shows the performance (in terms of F-Measure) of BiSLM for different combinations of values of η and ω parameters in the analysis of methylation profiles generated from synthetic epialleles at 50x of sequencing coverage.

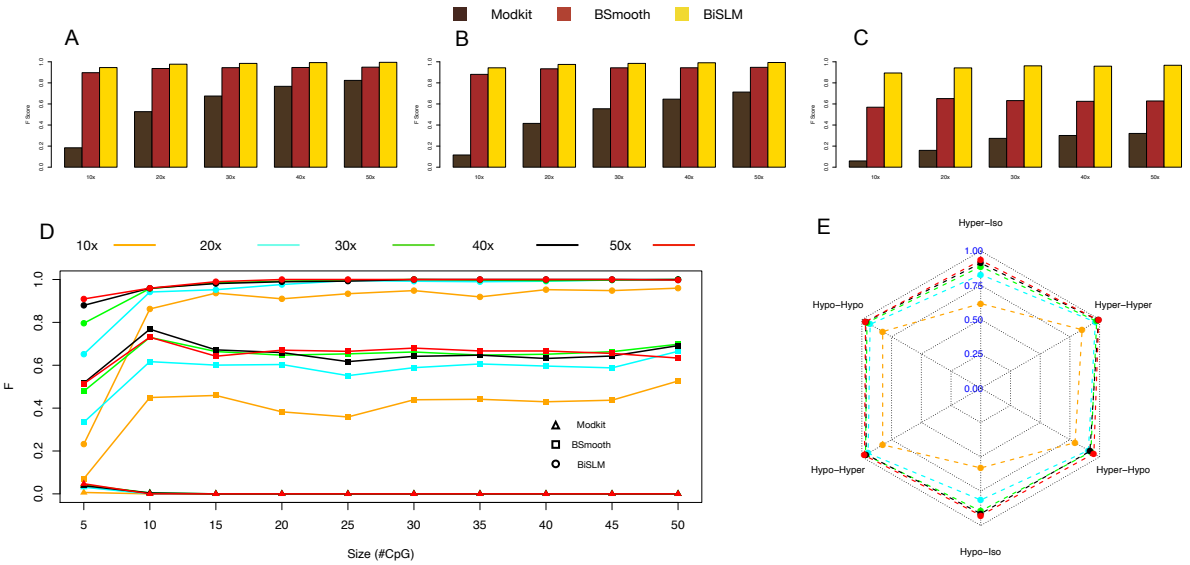

Figure S8: Performance of BiSLM, Modkit and BSmooth on the analysis of synthetic methylation profiles. Panels (a–c) present the F-measure values (harmonic mean of precision and recall) obtained by BiSLM, Modkit, and BSmooth for the detection of synthetic DMRs, assessed at varying reciprocal overlap thresholds: 0.1 in panel (A), 0.5 in panel (B), and 0.9 in panel (C). Panel (D) reports the F-measure obtained by BiSLM, Modkit and BSmooth in the detection of synthetic DMRs of different size as a function of sequencing coverage. The radarplot of panel (E) shows the correct classification rate (CCR) obtained by bivariate SLM in correctly classifying different classes of epiallelic shifts for different sequencing coverages.

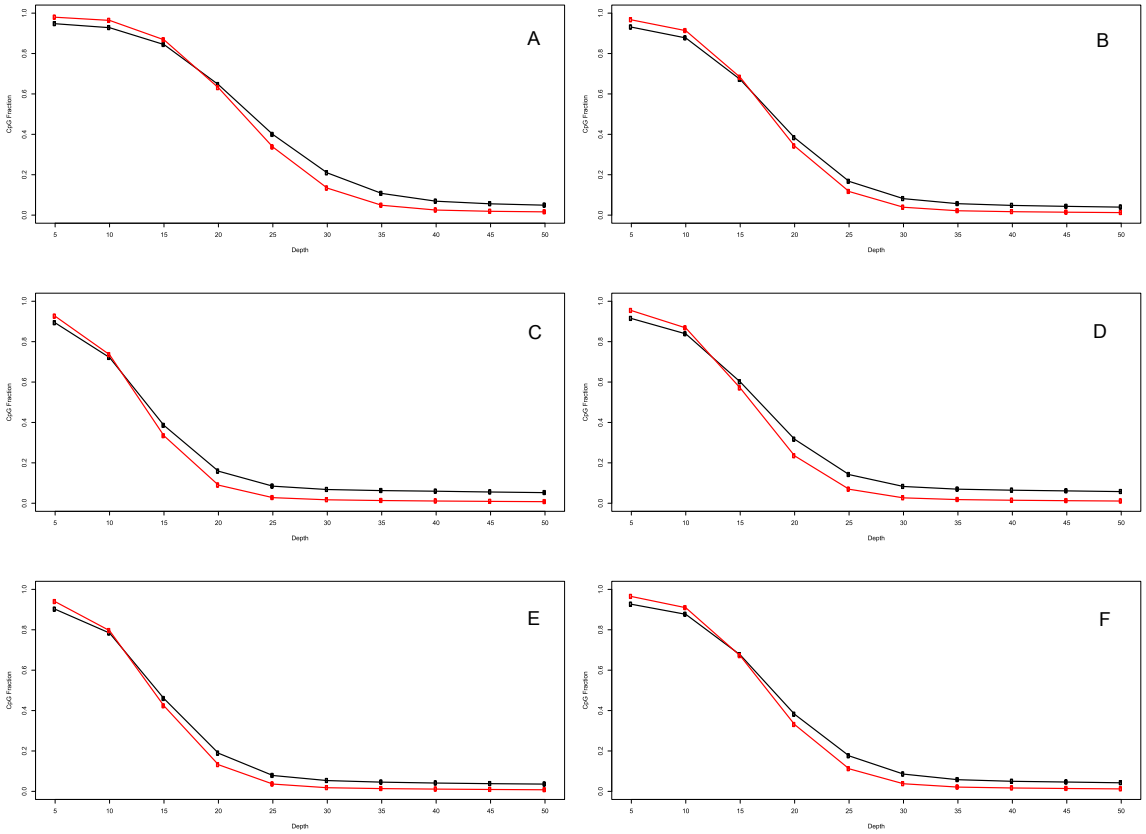

Figure S9: The figure shows the fraction of covered CpGs as a function of sequencing depth across the six AML samples. Black lines represent coverage across all genomic regions, while red lines correspond to repeat regions. Results are shown for UD5T (A), UD5R (B), UD10T (C), UD10R (D), AML2T (E), and AML2R (F).

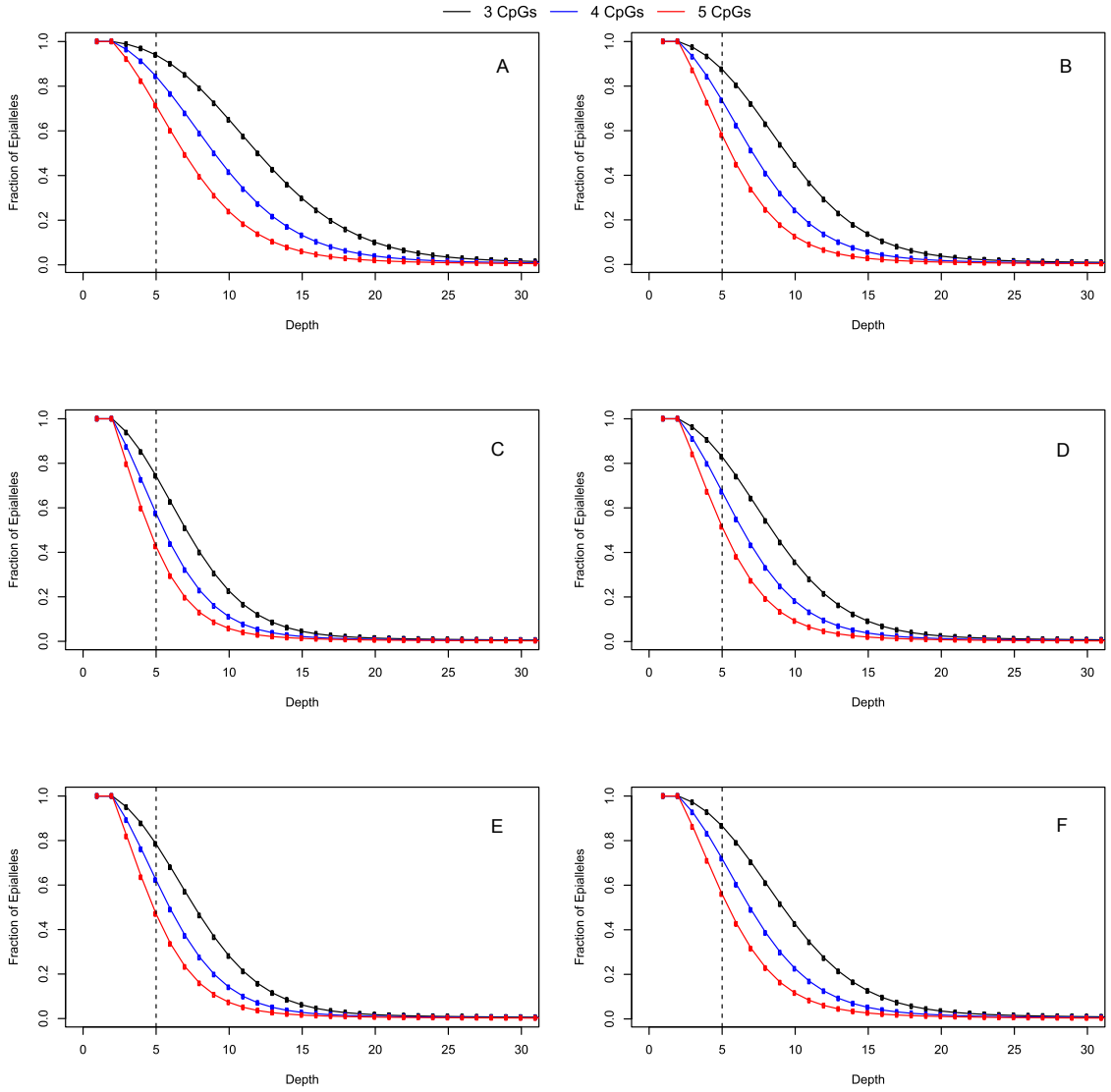

Figure S10: The figure shows the fraction of epialleles as a function of sequencing coverage (depth) for the six AML samples for different values of b (b=3 CpGs, b=4 CpGs and b=5 CpGs). Results are shown for UD5 Tum (A), UD5 Rel (B), UD10 Tum (C), UD10 Rel (D), AML2 Tum (E), and AML2 Rel (F). Vertical dotted line indicates 5x coverage.

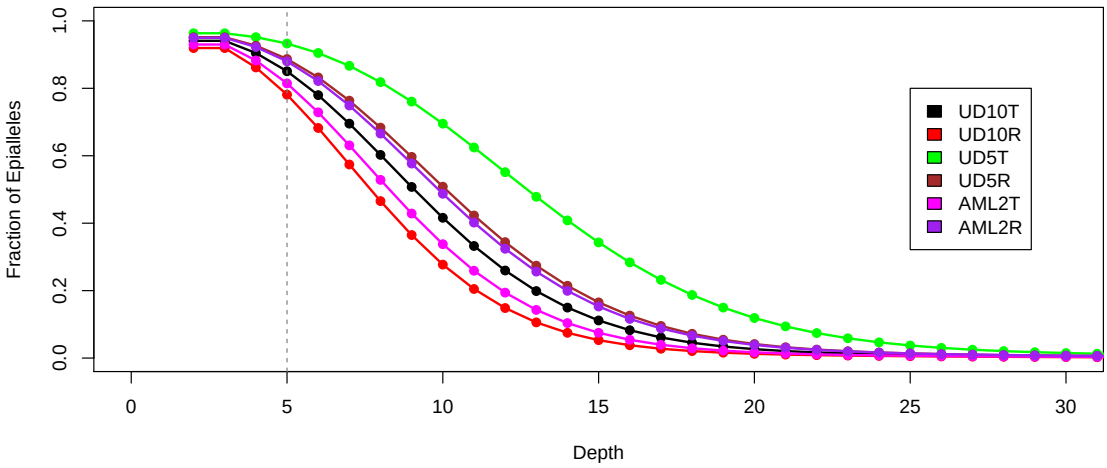

Figure S11: The figure shows the fraction of epialleles as a function of sequencing coverage (depth) for the six AML samples (for b=3 CpGs). Vertical dotted line indicates 5x coverage.

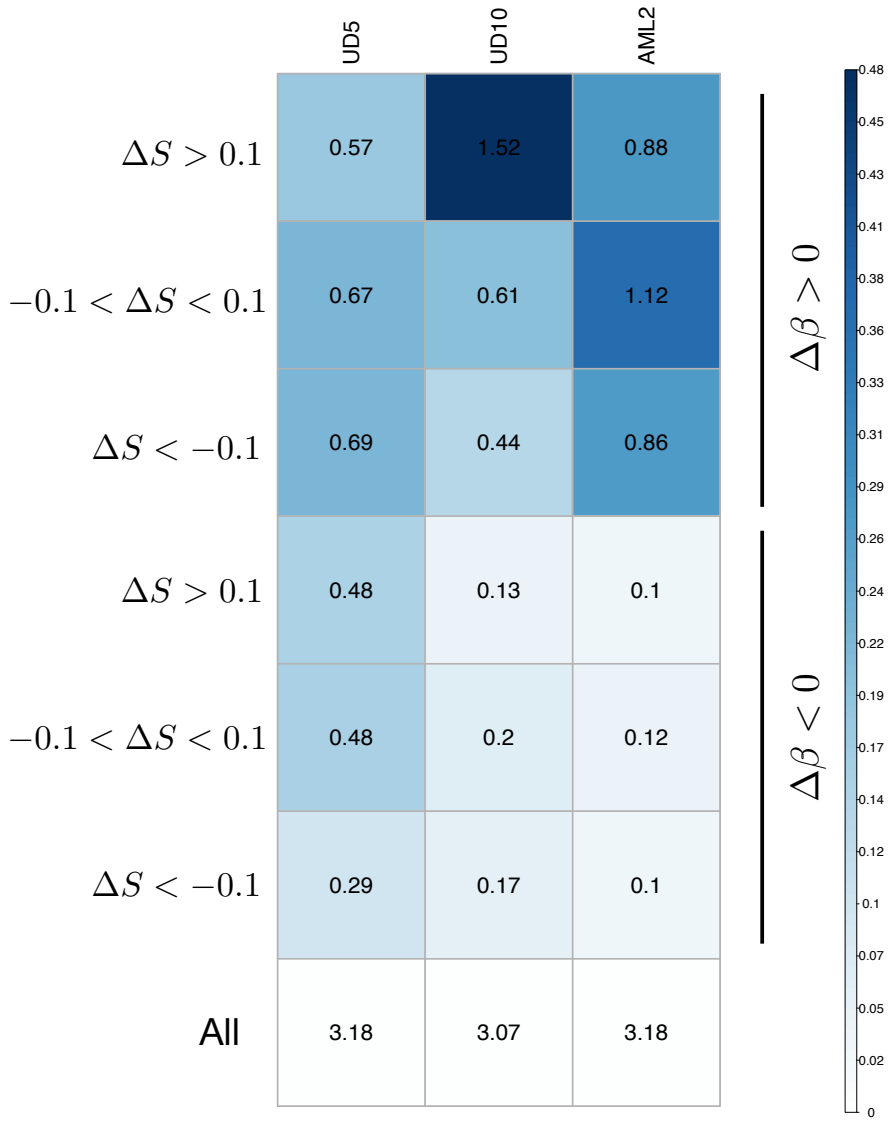

Figure S12: The figure shows the cumulative size of DMRs (in Mb), overall and for different ΔS and $\Delta\beta$ categories for the three AML sample pairs. While single DMRs show a similar size distribution across the six categories, the cumulative size is significantly higher for hyper-methylated regions ($\Delta\beta > 0.2$) compared to hypo-methylated regions ($\Delta\beta <$ -0.2)

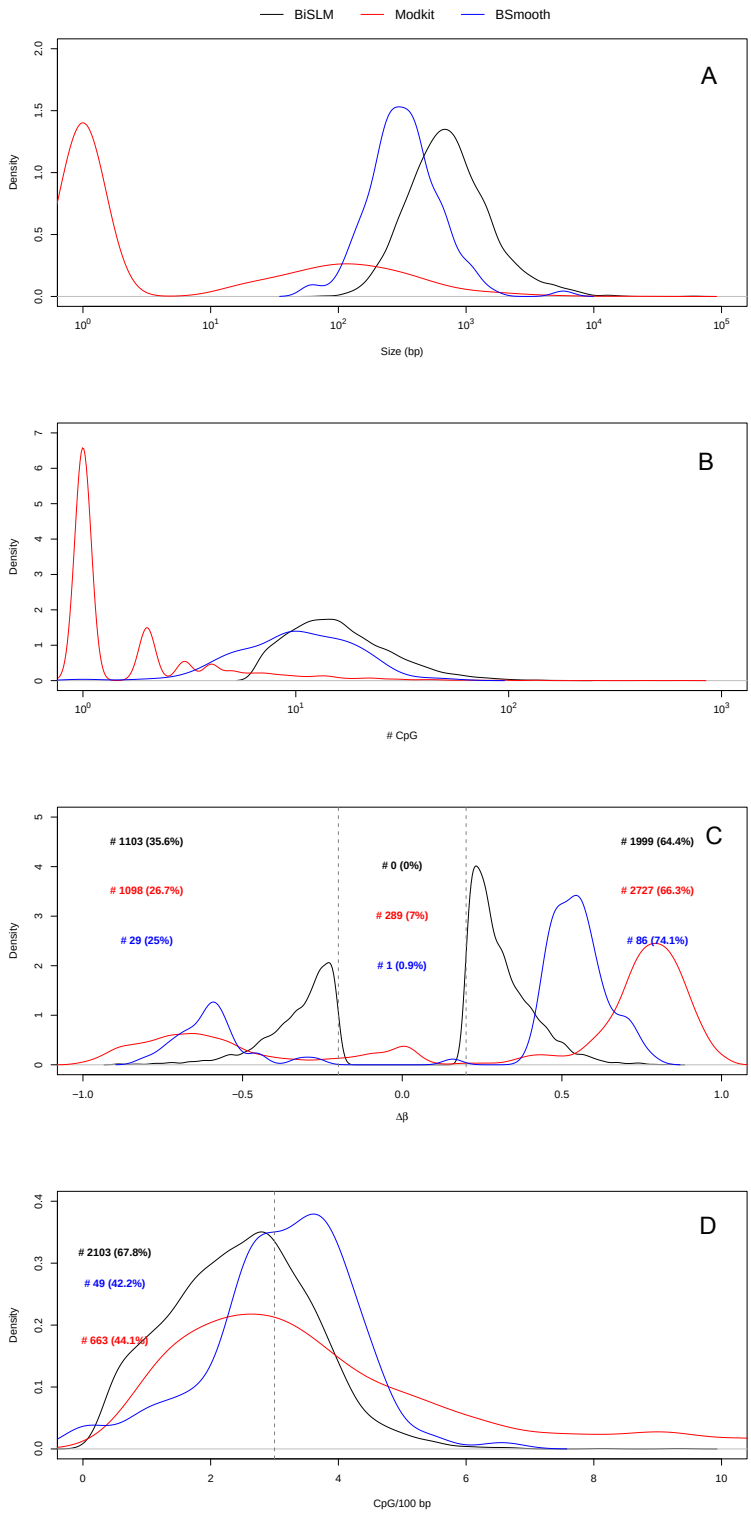

Figure S13: The figure reports the results obtained by BiSLM, Modkit, and BSmooth in the analysis of the UD5 sample pairs. Panel (A) shows the distribution of DMR sizes in base pairs, panel (B) shows the distribution in terms of number of CpGs, panel (C) illustrates the CpG density, and panel (D) displays the distribution of $\Delta\beta$ values. In panel D, the total number of DMRs identified by each method is also shown, categorized as hypo-methylated ($\Delta\beta < -0.2$), hyper-methylated ($\Delta\beta > 0.2$), or not differentially methylated ($-0.2 < \Delta\beta < 0.2$). Panel C further reports the fraction and absolute number of DMRs with CpG density lower than 3 (indicated by the vertical dotted line, corresponding to low-density CpG regions).

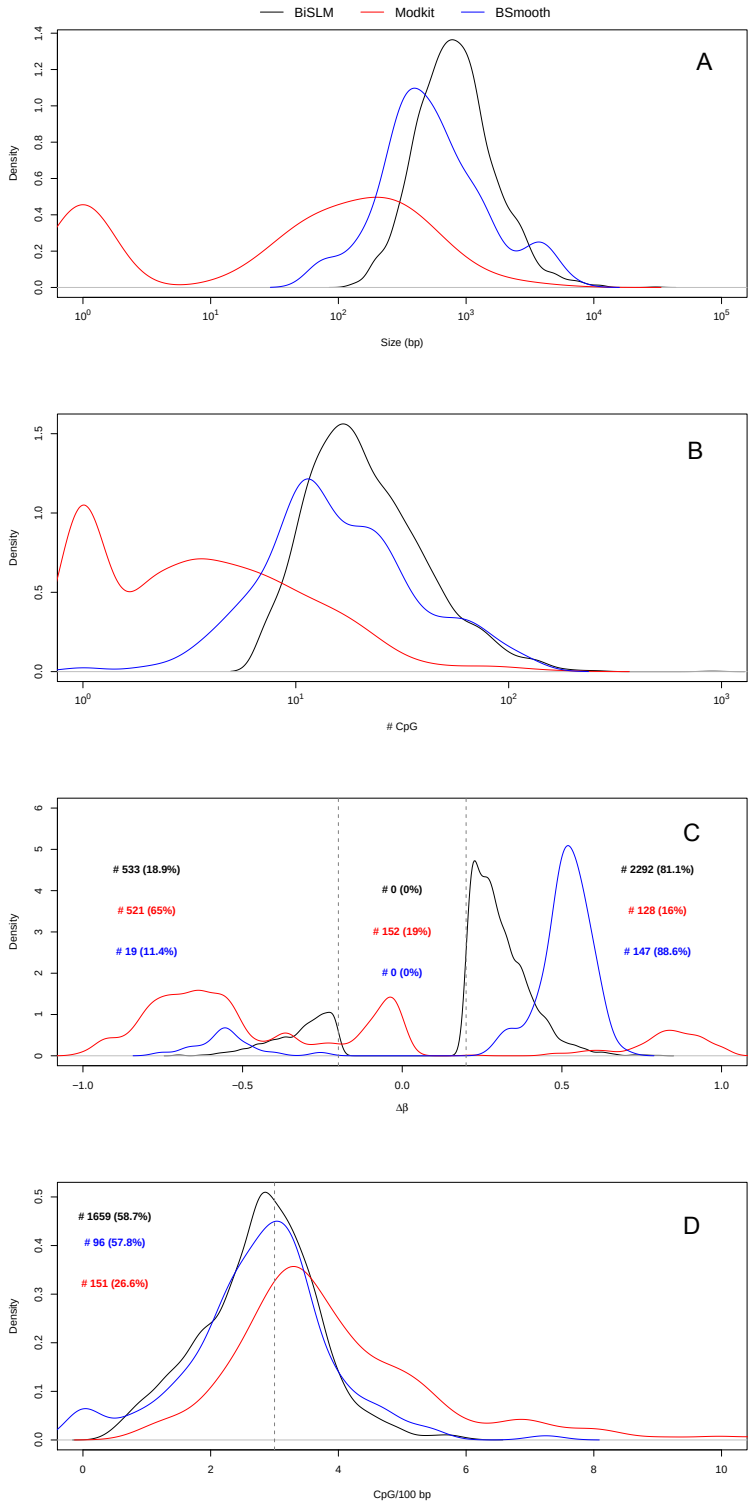

Figure S14: The figure reports the results obtained by BiSLM, Modkit, and BSmooth in the analysis of the UD10 sample pairs. Panel (A) shows the distribution of DMR sizes in base pairs, panel (B) shows the distribution in terms of number of CpGs, panel (C) illustrates the CpG density, and panel (D) displays the distribution of $\Delta\beta$ values. In panel D, the total number of DMRs identified by each method is also shown, categorized as hypo-methylated ($\Delta\beta < -0.2$), hyper-methylated ($\Delta\beta > 0.2$), or not differentially methylated ($-0.2 < \Delta\beta < 0.2$). Panel C further reports the fraction and absolute number of DMRs with CpG density lower than 3 (indicated by the vertical dotted line, corresponding to low-density CpG regions).

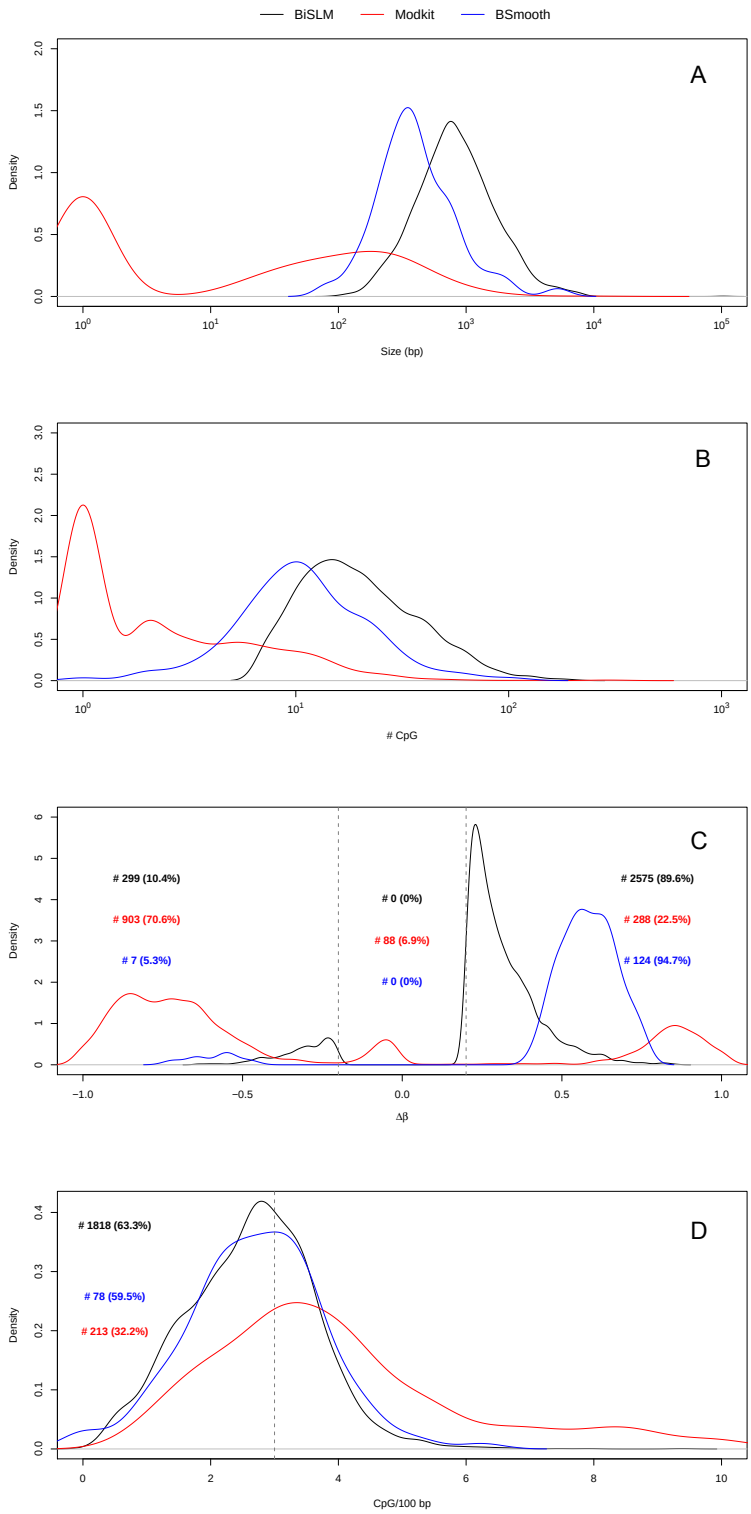

Figure S15: The figure reports the results obtained by BiSLM, Modkit, and BSmooth in the analysis of the AML2 sample pairs. Panel (A) shows the distribution of DMR sizes in base pairs, panel (B) shows the distribution in terms of number of CpGs, panel (C) illustrates the CpG density, and panel (D) displays the distribution of $\Delta\beta$ values. In panel D, the total number of DMRs identified by each method is also shown, categorized as hypo-methylated ($\Delta\beta < -0.2$), hyper-methylated ($\Delta\beta > 0.2$), or not differentially methylated ($-0.2 < \Delta\beta < 0.2$). Panel C further reports the fraction and absolute number of DMRs with CpG density lower than 3 (indicated by the vertical dotted line, corresponding to low-density CpG regions).

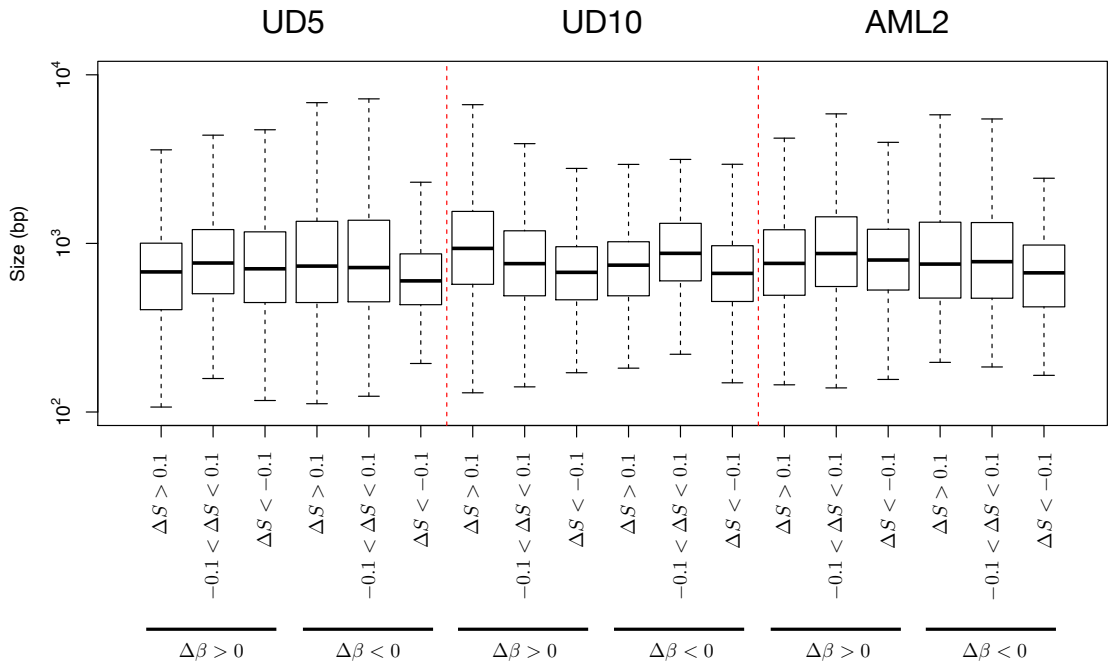

Figure S16: The figure shows DMRs size distribution for different ΔS and $\Delta\beta$ categories across the three AML sample pairs.

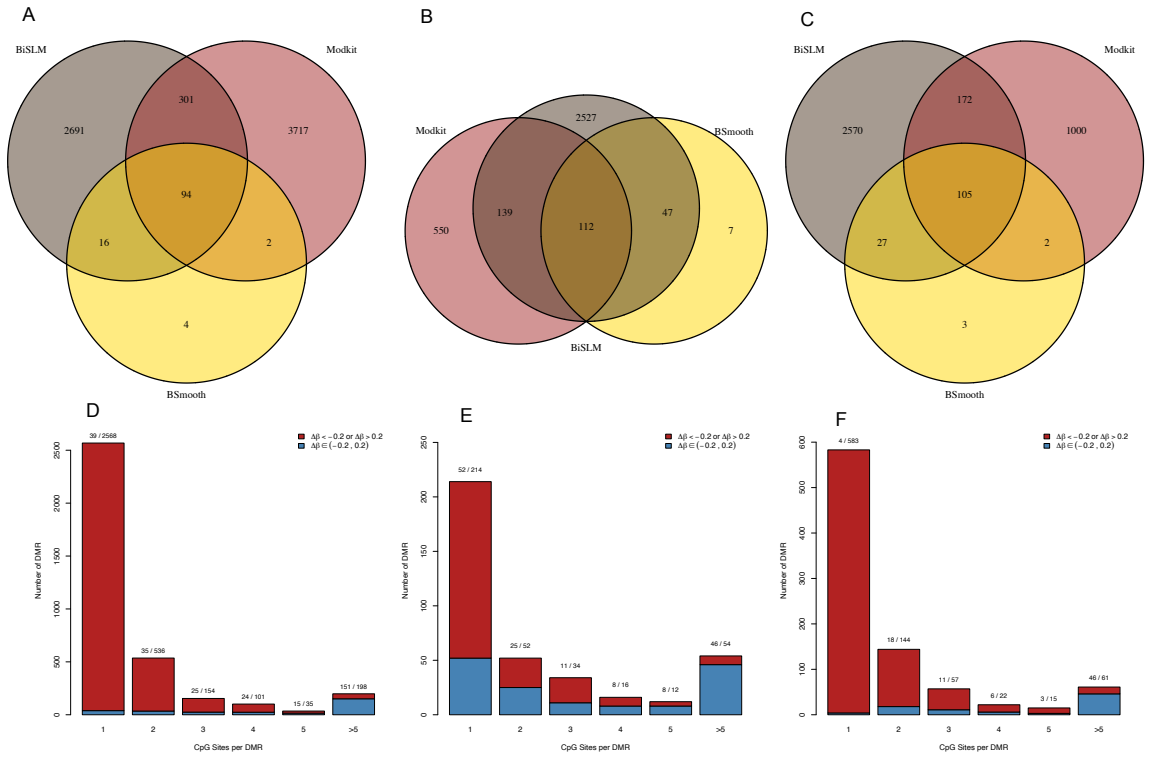

Figure S17: The figure shows the overlap between the DMRs identified by BiSLM, Modkit, and BSmooth in the analysis of the three sample pairs (panel A: UD5, panel B: UD10, panel C: AML2). Panels D-F display barplots illustrating the distribution of the number of CpGs within the DMRs identified by Modkit (panel D: UD5, panel E: UD10, panel F: AML2). Results are reported separately for DMRs classified as non-differentially methylated ($-0.2 < \Delta\beta < 0.2$) and differentially methylated ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$).

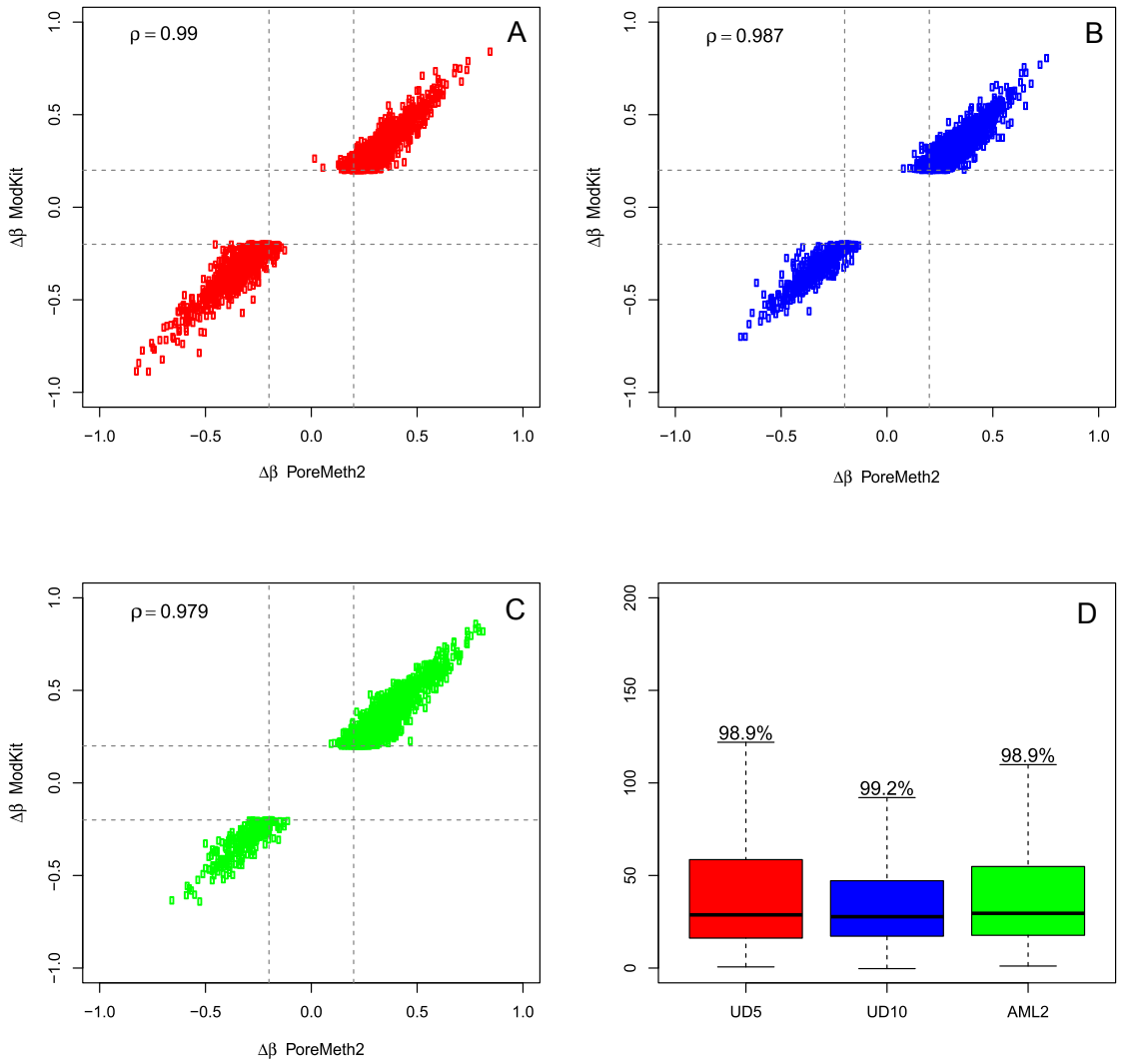

Figure S18: Panel A-C show the correlation between the $\Delta\beta$ values calculated by BiSLM and Modkit for the DMRs exclusively identified by our tool (A for UD5, B for UD10 and C for AML2). The boxplot of panel D shows the distribution of the Modkit LR score calculated for all the DMRs exclusively identified by our tool.

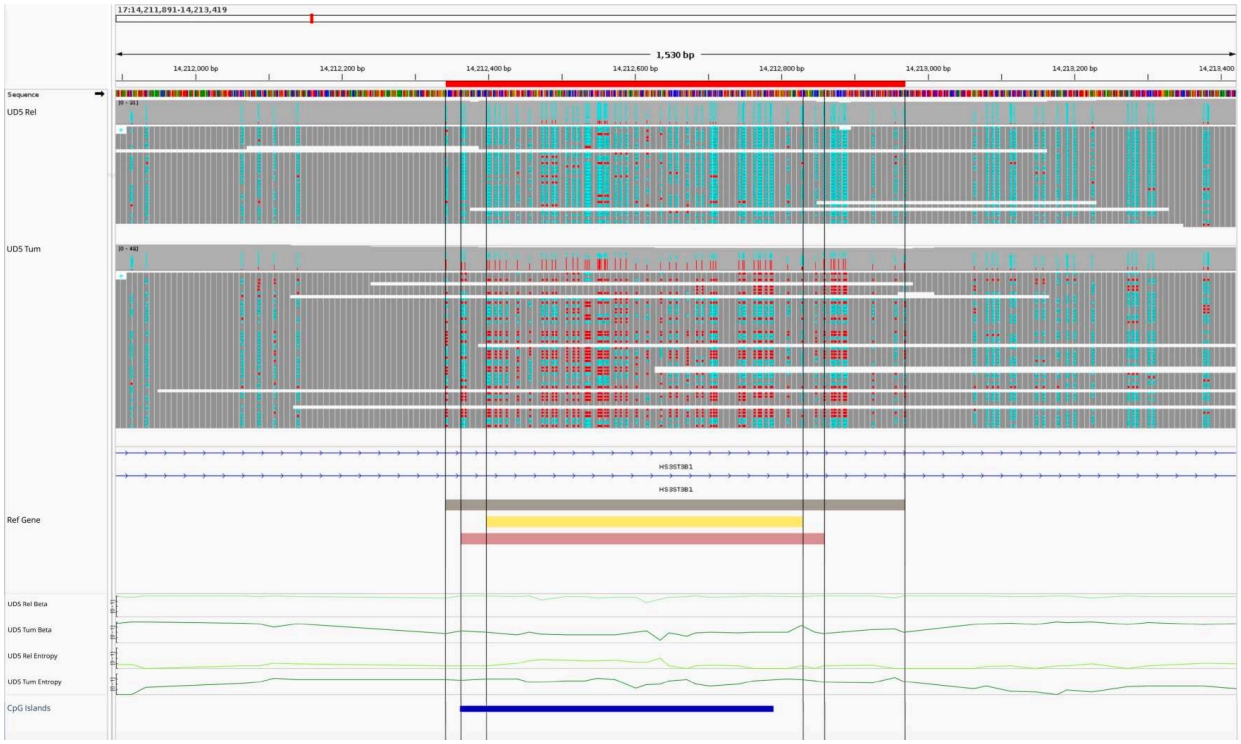

Figure S19: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the UD5 sample pair in a high CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown, as well as CpG Island overlap (blue).

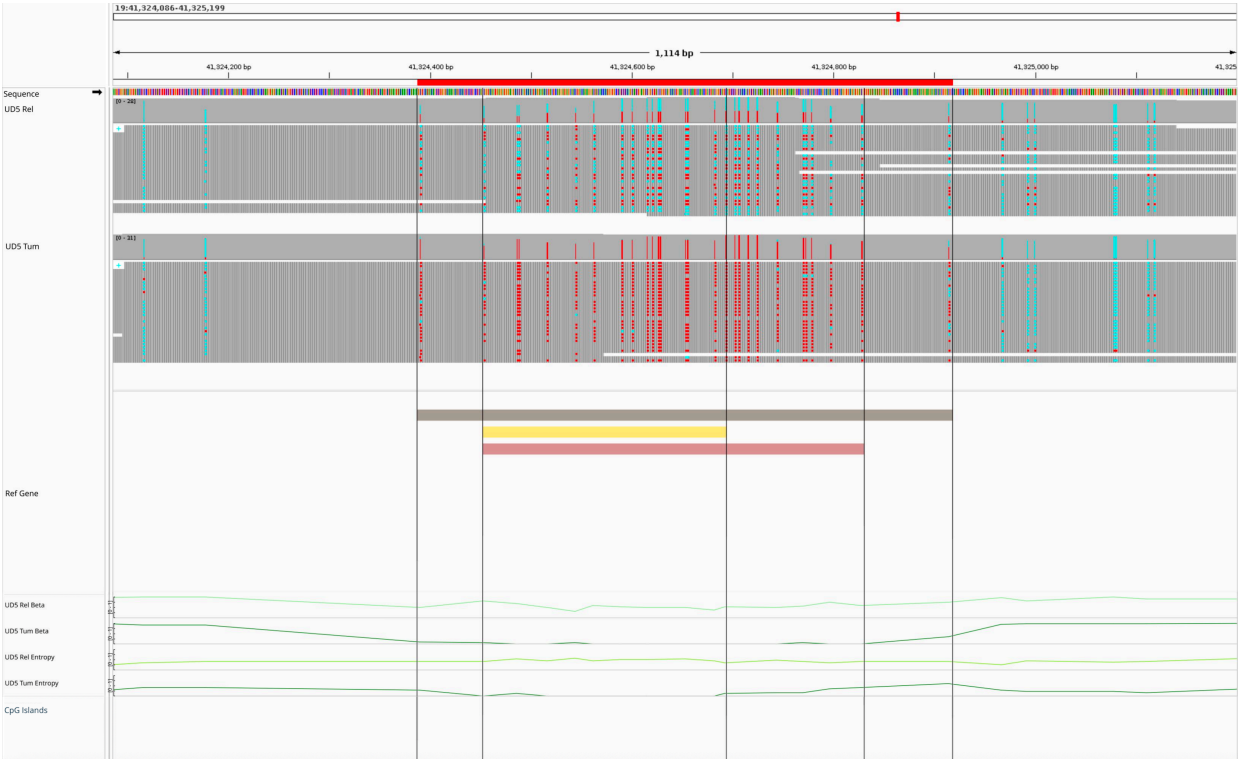

Figure S20: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the UD5 sample pair in a low CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown.

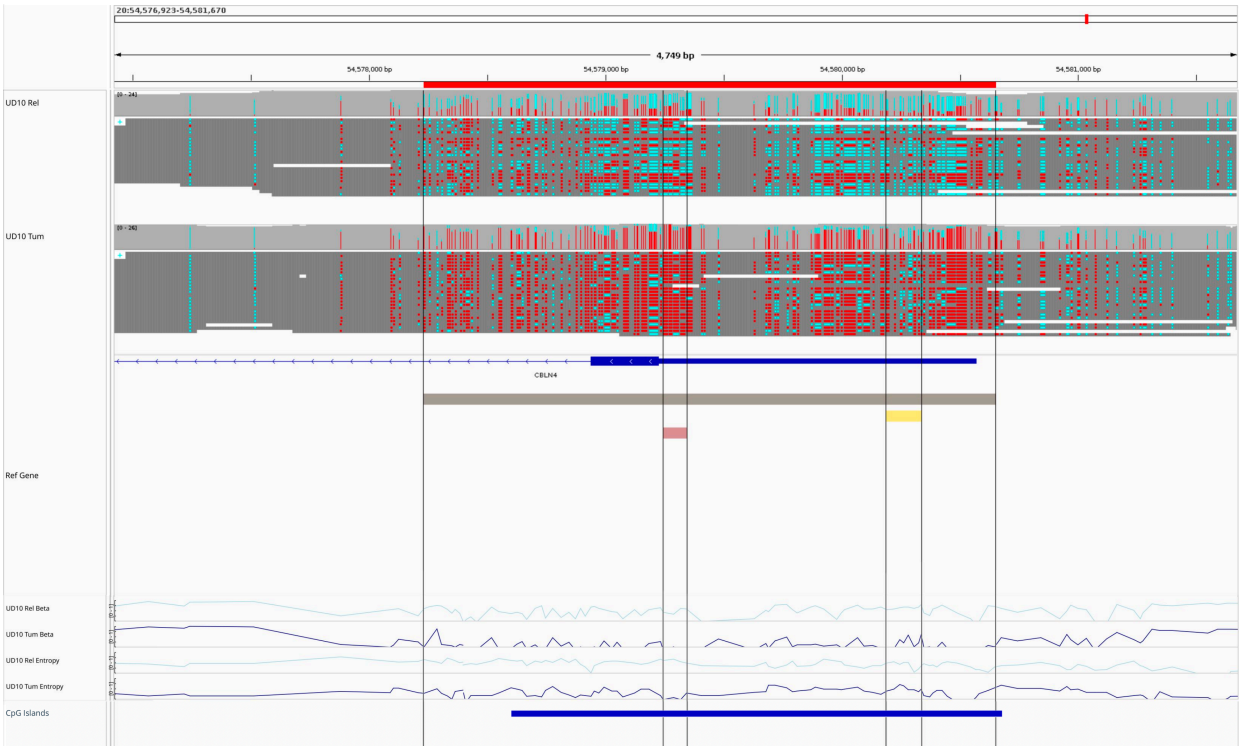

Figure S21: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the UD10 sample pair in a high CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown, as well as CpG Island overlap (blue).

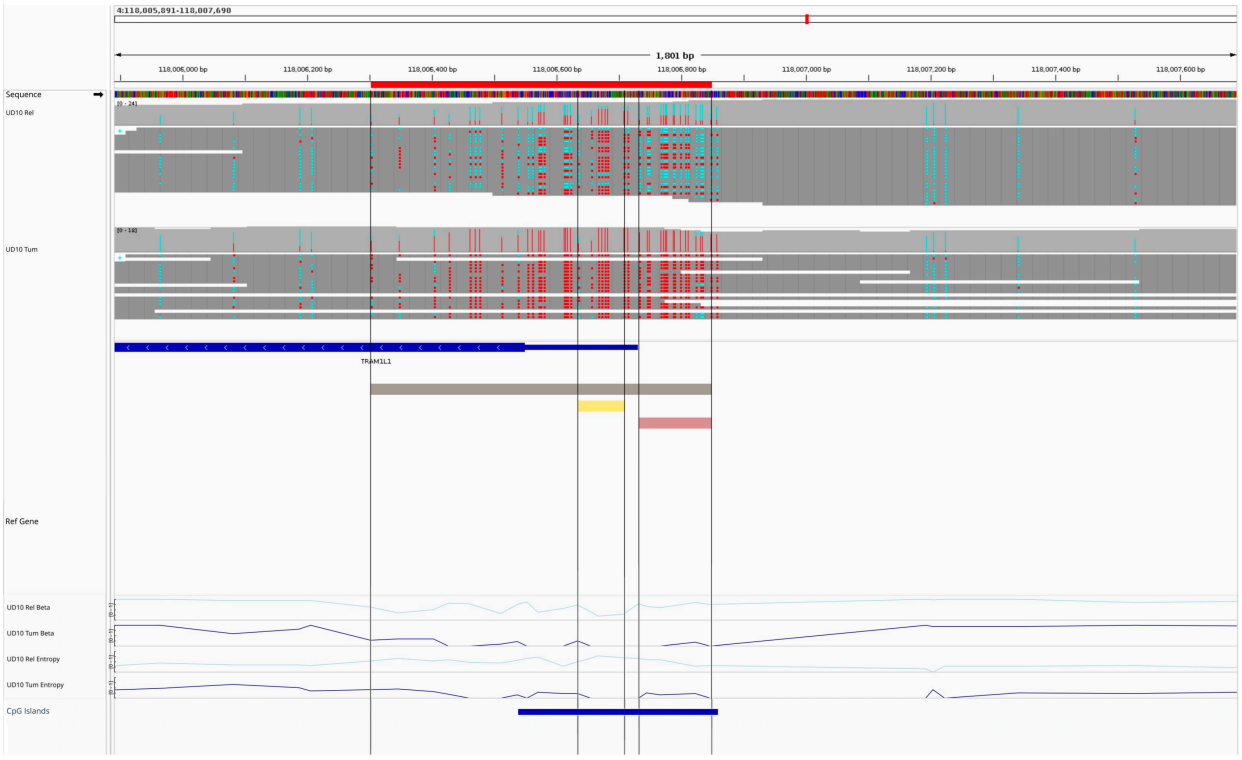

Figure S22: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the UD10 sample pair in a high CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown, as well as CpG Island overlap (blue).

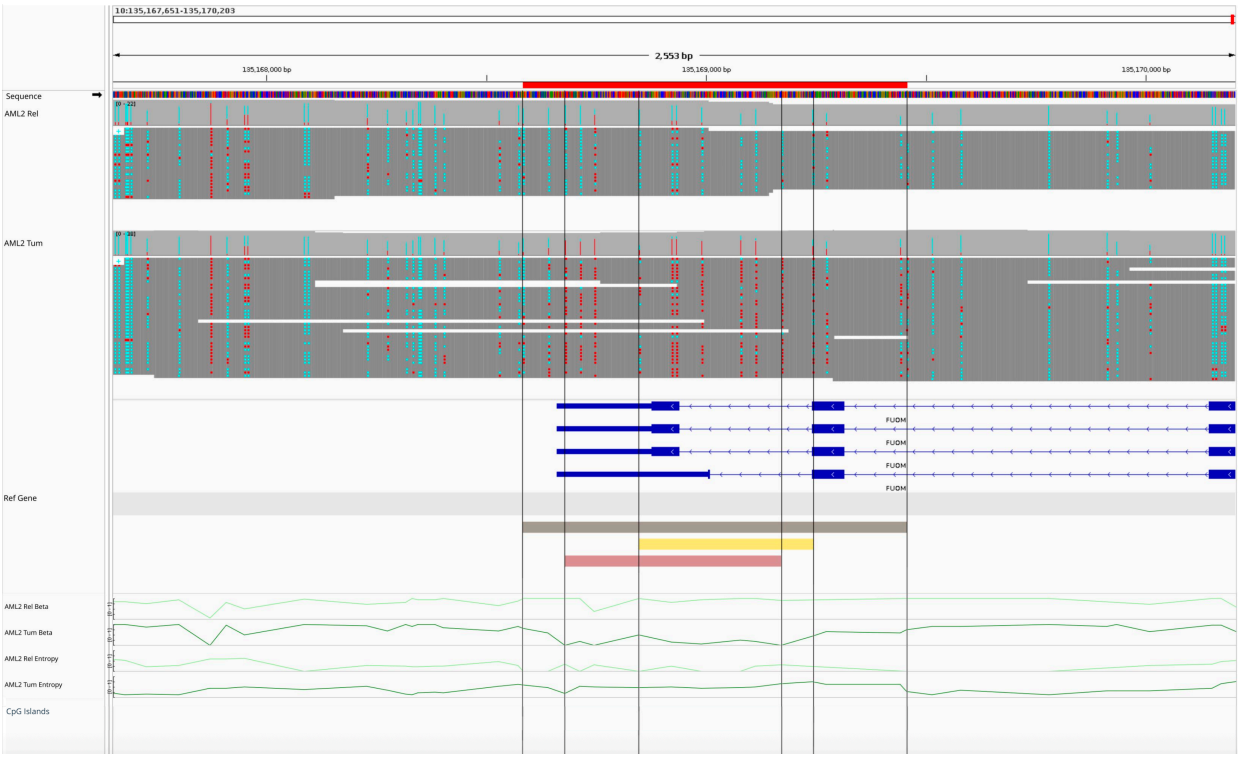

Figure S23: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the AML2 sample pair in a low CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown.

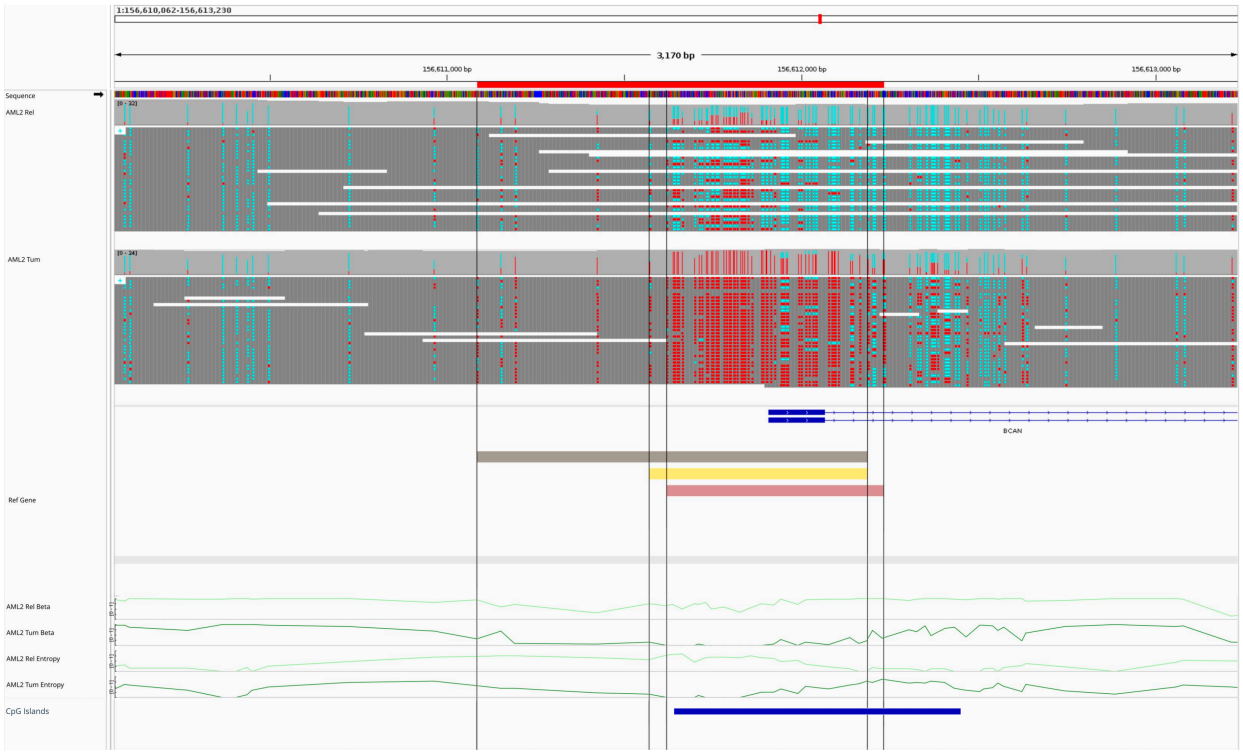

Figure S24: The figures display resulting plots from Integrative Genomics Viewer (IGV) of DRMs identified by both BSmooth, Poremeth2 and Modkit in the AML2 sample pair in a high CpG-density region with a reciprocal overlap 50% between the three algorithms. Brown tracks represent high-confidence DRMs detected by Poremeth2, yellow tracks indicate corresponding regions identified by BSmooth and red tracks those identified by Modkit. β and S values across genomic positions in both Tumor (dark tracks) and Relapse (light tracks) samples are also shown, as well as CpG Island overlap (blue).

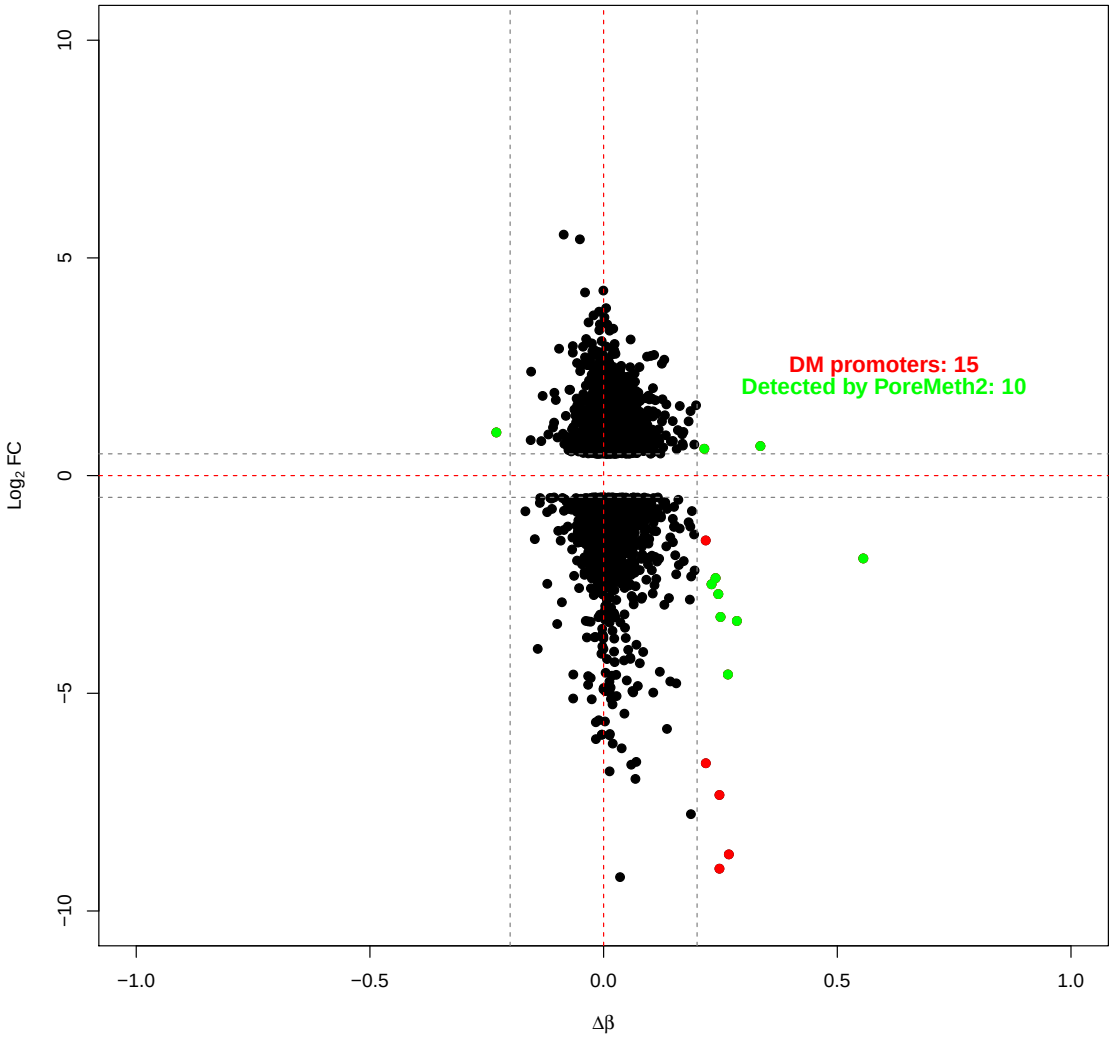

Figure S25: The figure shows the average $\Delta\beta$ values across the promoter regions of all differentially expressed genes (DEGs) identified in the UD5 sample pairs. Green dots represent DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) where a DMR was also detected by PoreMeth2 within the promoter region. Red dots indicate DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) but without a corresponding DMR identified by PoreMeth2.

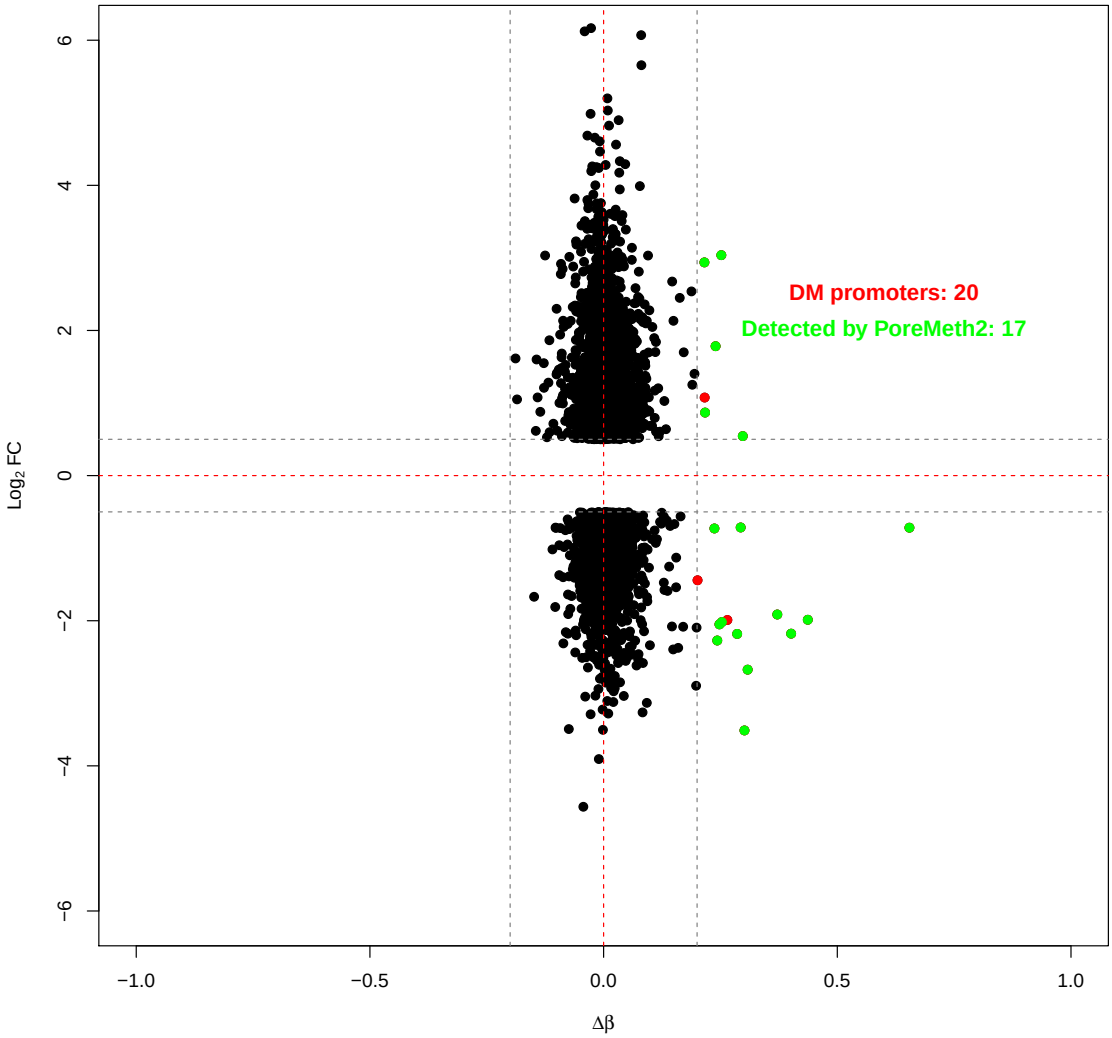

Figure S26: The figure shows the average $\Delta\beta$ values across the promoter regions of all differentially expressed genes (DEGs) identified in the UD10 sample pairs. Green dots represent DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) where a DMR was also detected by PoreMeth2 within the promoter region. Red dots indicate DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) but without a corresponding DMR identified by PoreMeth2.

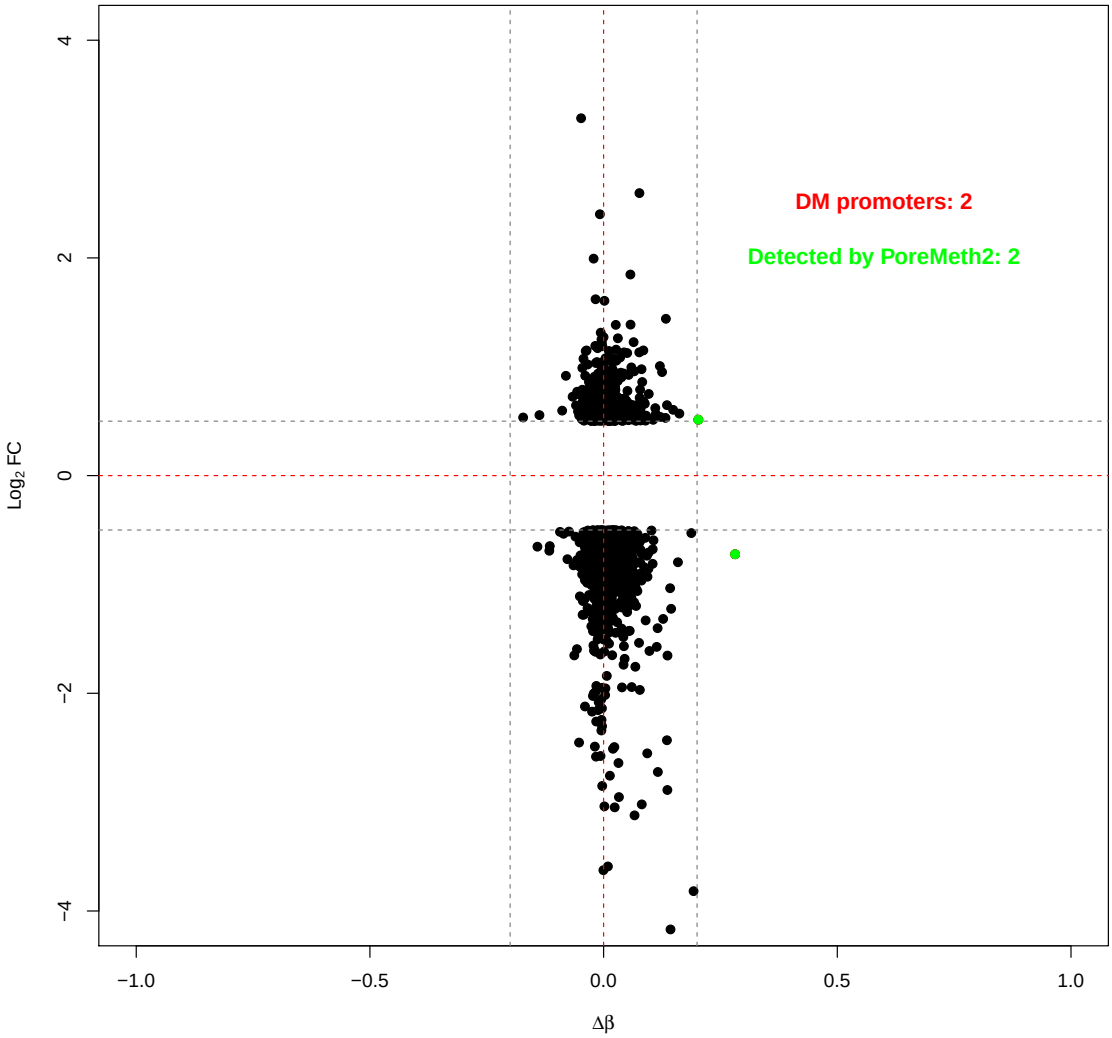

Figure S27: The figure shows the average $\Delta\beta$ values across the promoter regions of all differentially expressed genes (DEGs) identified in the AML2 sample pairs. Green dots represent DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) where a DMR was also detected by PoreMeth2 within the promoter region. Red dots indicate DEGs with differentially methylated promoters ($\Delta\beta > 0.2$ or $\Delta\beta < -0.2$) but without a corresponding DMR identified by PoreMeth2.

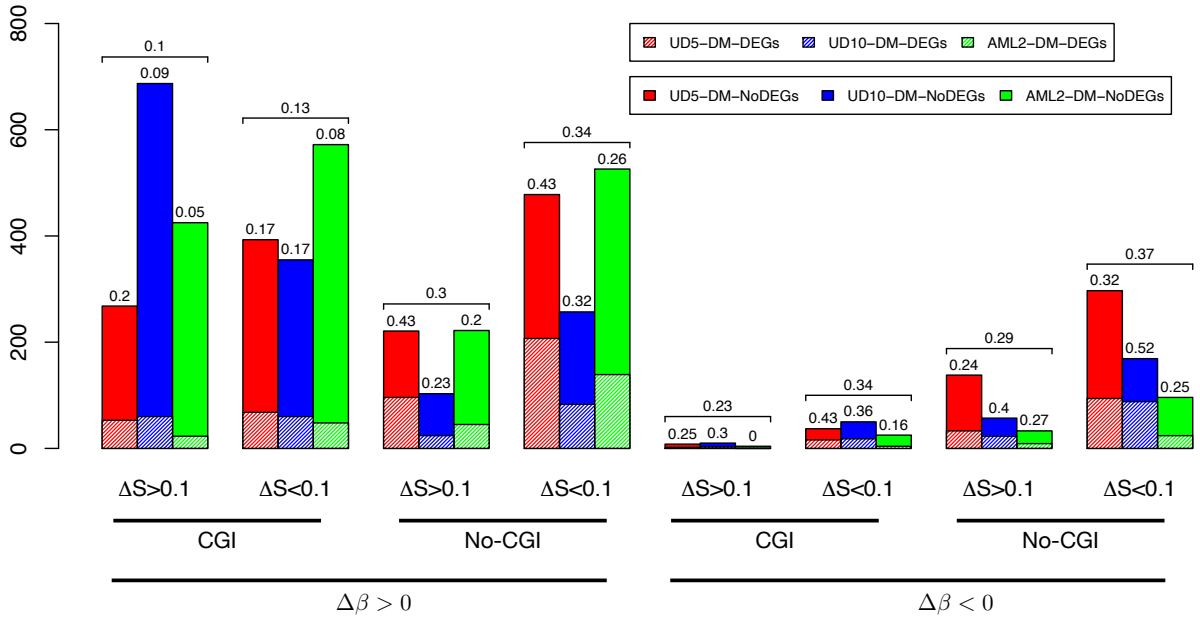

Figure S28: Figure shows the proportion of DMGs that are also DEGs for different ΔS and $\Delta\beta$ categories for the three AML sample pairs. Data are reported for DMRs overlapping CpG islands (CGI) and outside CpG islands (NoCGI). Textured bars show the number of DM-DEGs.

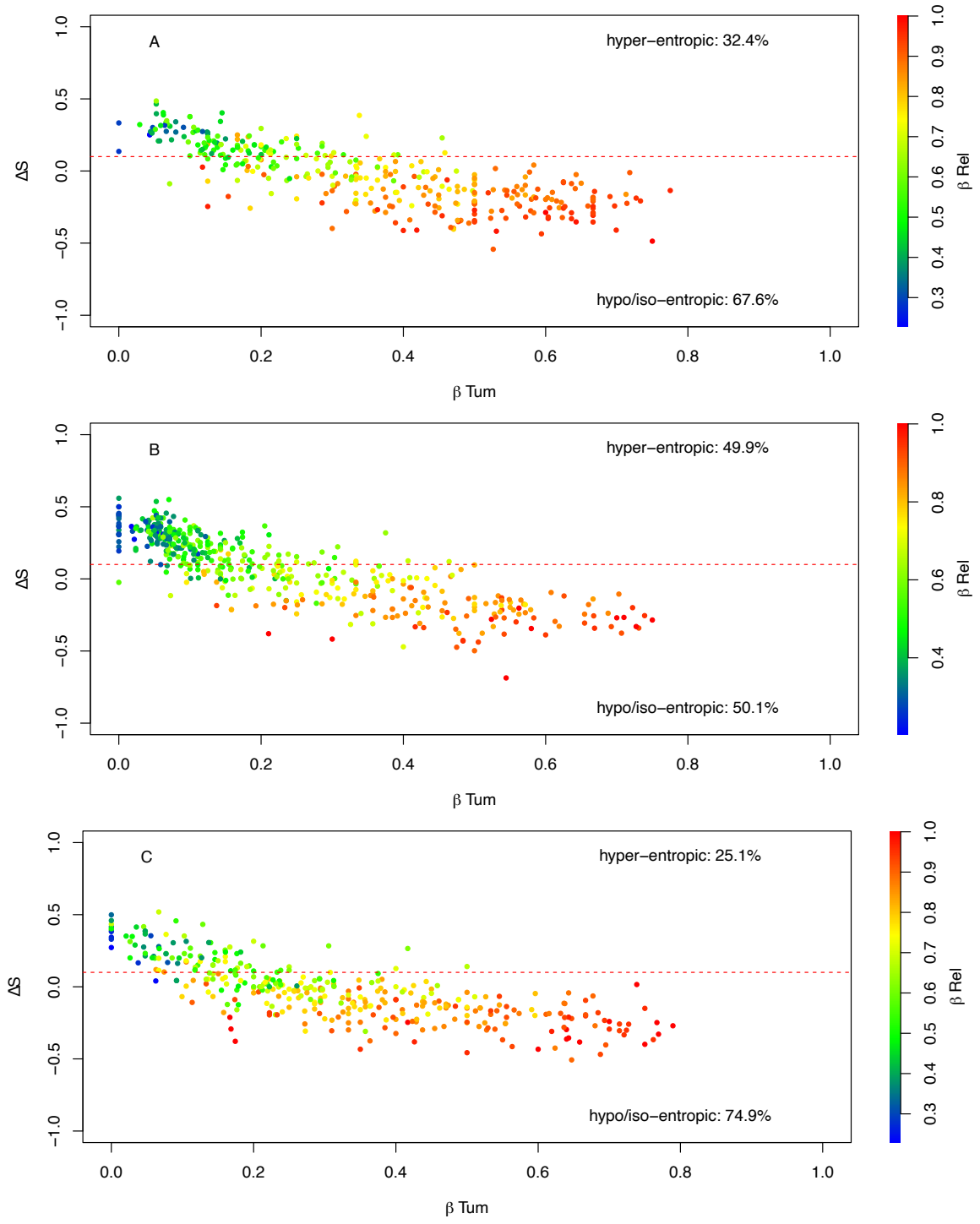

Figure S29: The scatterplots display epiallelic entropy changes (ΔS) as a function of baseline methylation level (β Tum) for DM-DEGs with hypermethylated DMRs in UD5 (A), UD10 (B) and AML2 (C). Points are colored according to the final methylation level (β Rel, blue to red gradient). The dashed red line indicates the $\Delta S = 0.1$ threshold used to distinguish hyperentropic from hypo/isoentropic changes.

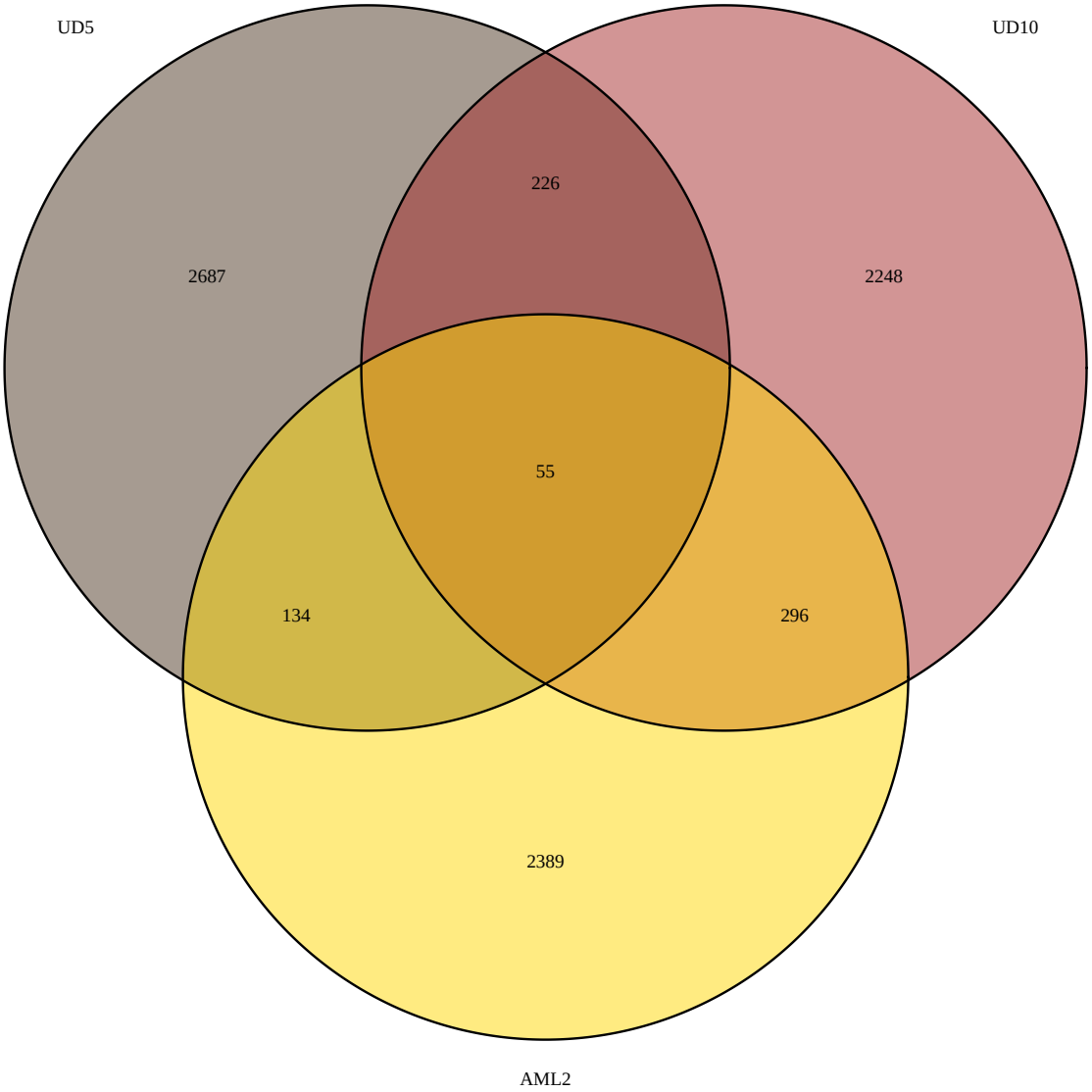

Figure S30: Venn diagram of the DMRs identified by BiSLM in the analysis of three AML sample pairs.

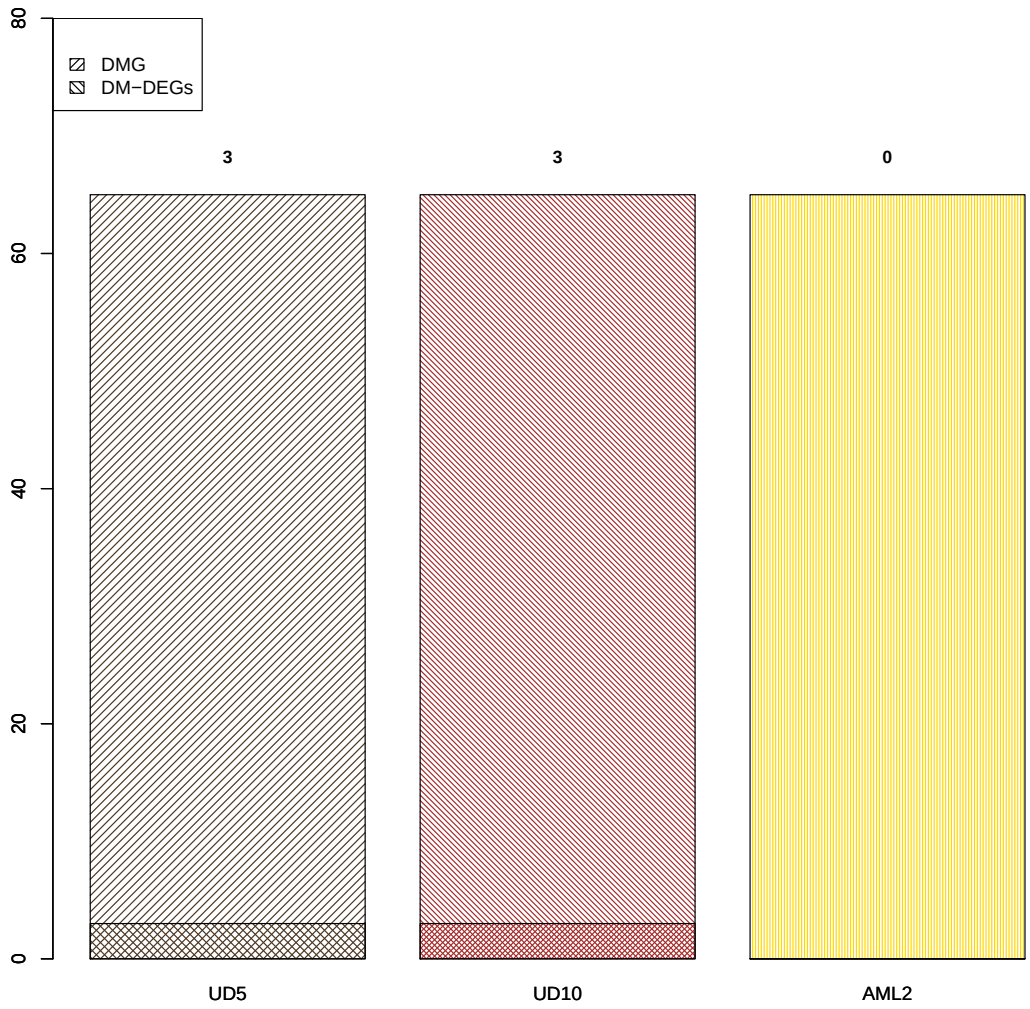

Figure S31: The figure shows the number of DMGs common to all three AML sample pairs and the corresponding number of DM-DEGs.

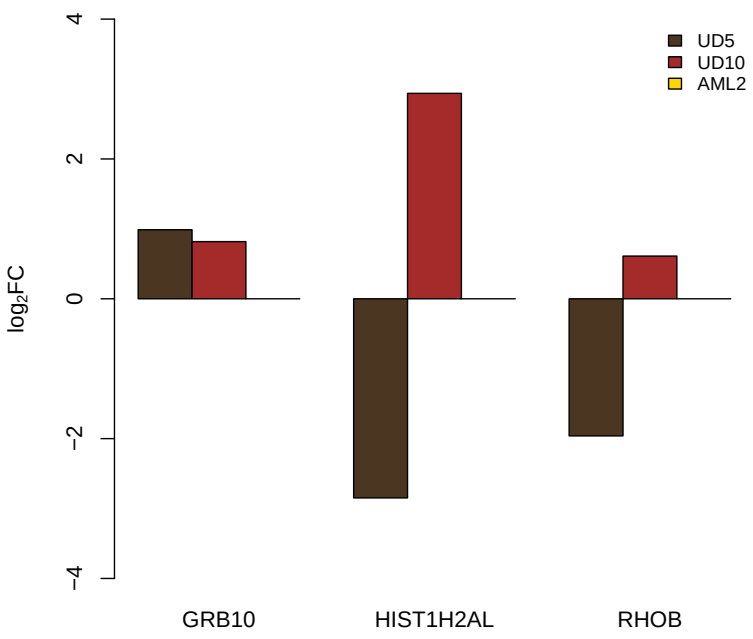

Figure S32: Expression levels of the three DM-DEGs common to all three AML sample pairs.

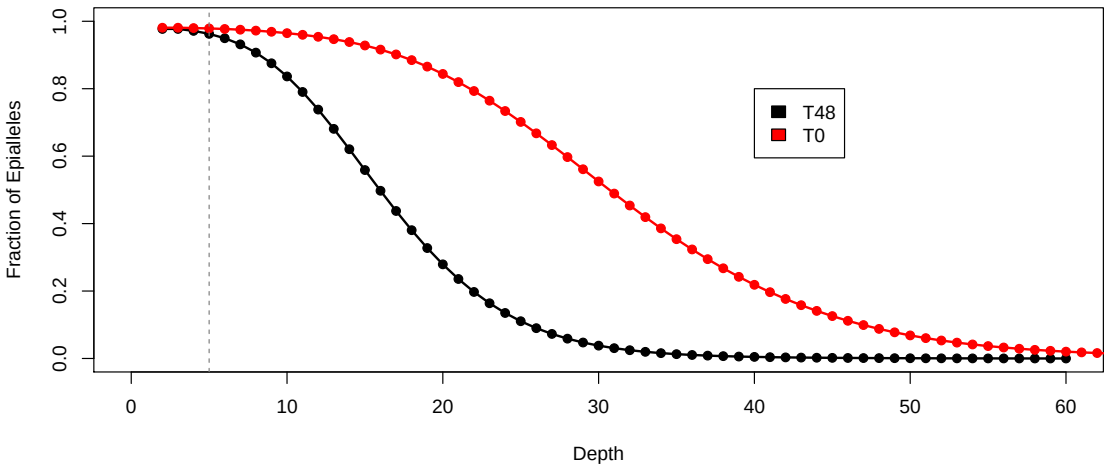

Figure S33: The figure shows the fraction of epialleles as a function of sequencing coverage (depth) for the HPGC dataset. Vertical dotted line indicates 5x coverage.

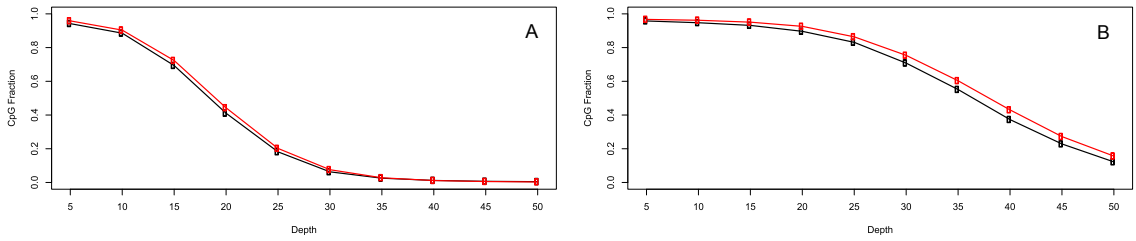

Figure S34: The figure shows the fraction of covered CpGs as a function of sequencing depth for the two HPGC samples. Black lines represent coverage across all genomic regions, while red lines correspond to repeat regions. Results are shown for T0 (A) and T48 (B).

	$\Delta\beta > 0$			$\Delta\beta < 0$			
T48	4.16	0.55	2.73	0	0.01	0.01	7.45
	$\Delta S > 0.1$	$-0.1 < \Delta S < 0.1$	$\Delta S < -0.1$	$\Delta S > 0.1$	$-0.1 < \Delta S < 0.1$	$\Delta S < -0.1$	All

Figure S35: The figure shows the cumulative size of DMRs (in Megabases), overall and for different ΔS and $\Delta\beta$ categories for the HPGC dataset. While single DMRs show a similar size distribution across the six categories, the cumulative size is significantly higher for hyper-methylated regions ($\Delta\beta > 0.2$) compared to hypo-methylated regions ($\Delta\beta <$ -0.2)

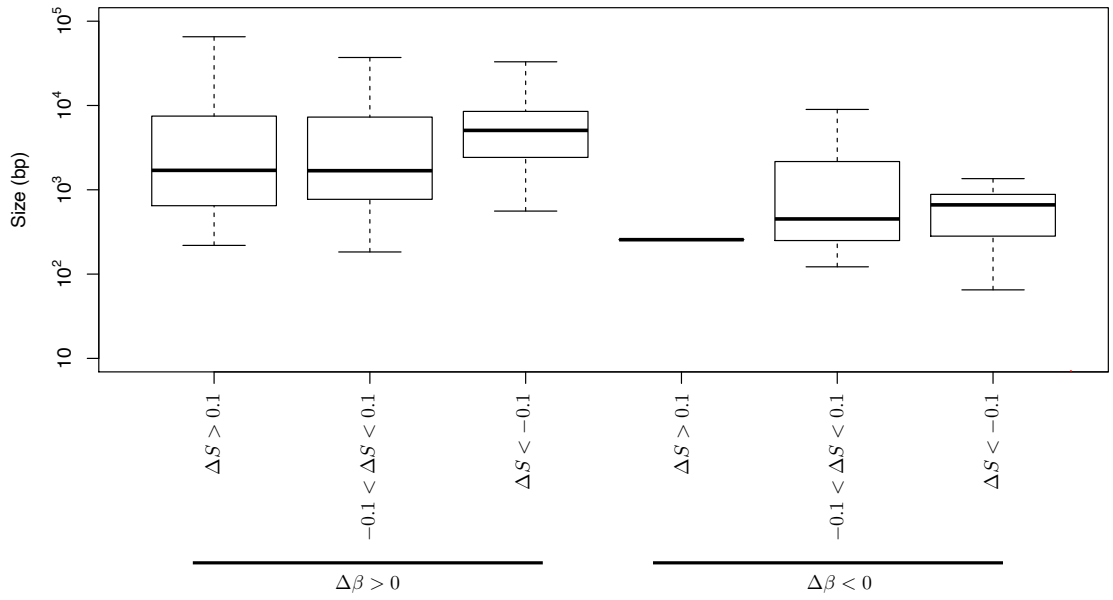

Figure S36: The figure shows DMRs size distribution for different ΔS and $\Delta\beta$ categories for the HPGC dataset.

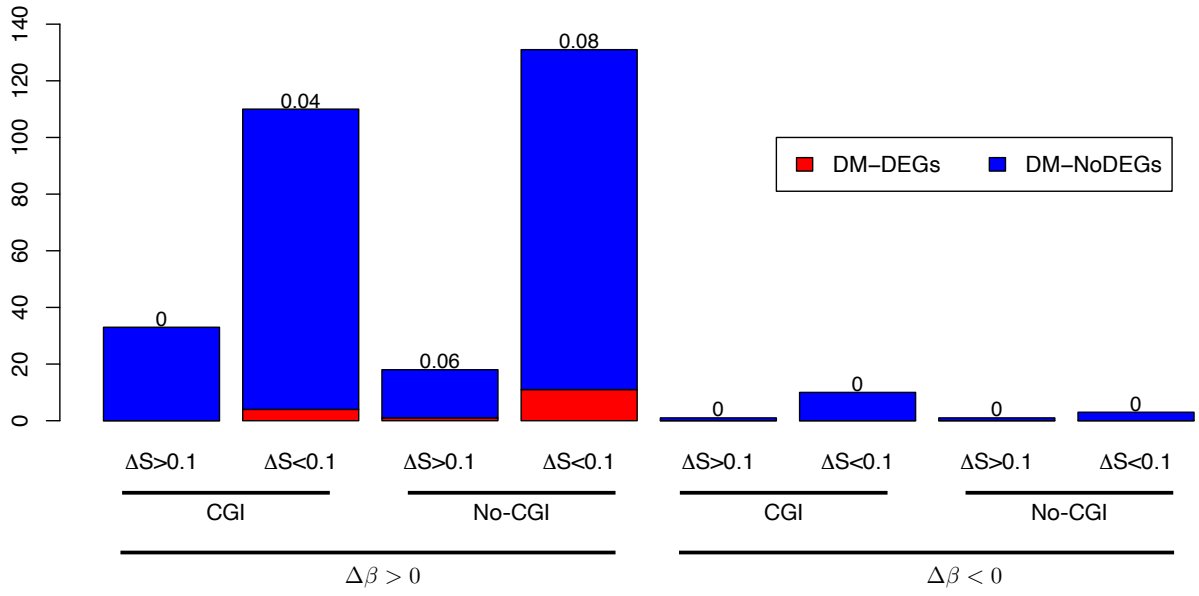

Figure S37: Figure shows the proportion of DMGs that are also DEGs for different ΔS and $\Delta\beta$ categories for the HPGC dataset. Data are reported for DMRs overlapping CpG islands (CGI) and outside CpG islands (NoCGI). Textured bars show the number of DM-DEGs.

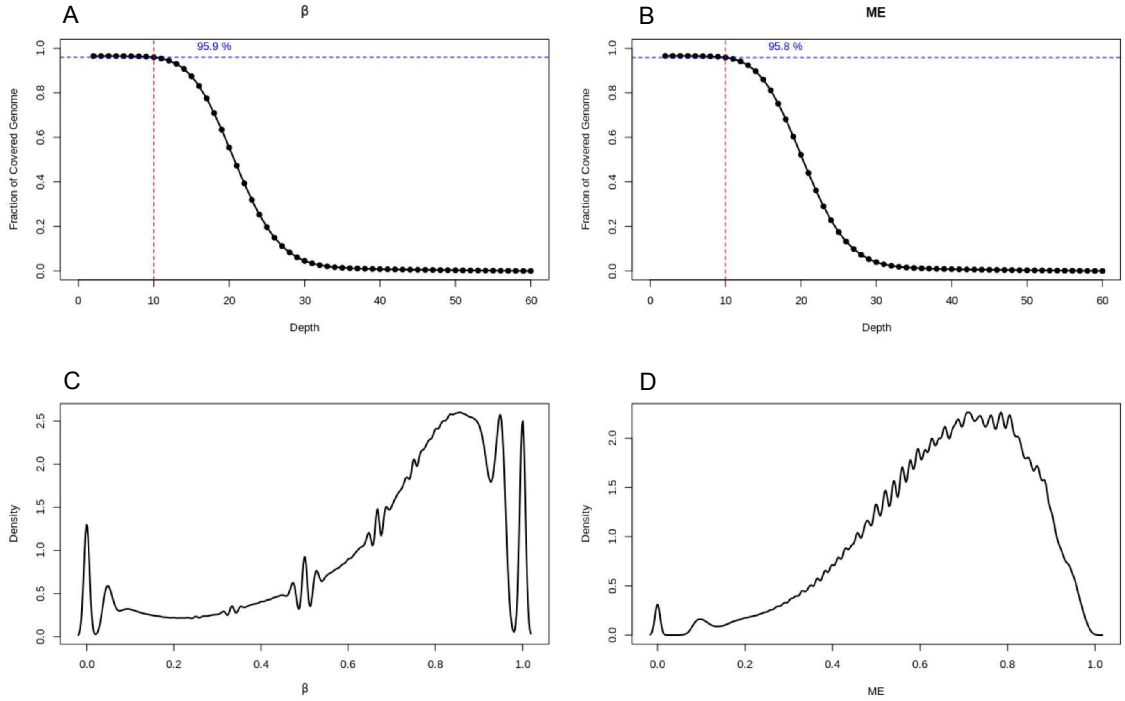

Figure S38: Figure shows an example of Quality Plot generated with PoreMeth2 function PoreMeth2SingleExpQualityPlot. The four quadrants respectively show: (A) the distribution of the number of reads (depth) used to calculate β in each CpG, (B) the distribution of the number of reads (depth) used to calculate S in each CpG, (C) the distribution of β values across the sample and (D) the distribution of S values across the sample. The vertical and horizontal lines in the first two quadrants highlight the fraction of CpGs covered by at least 10 reads.

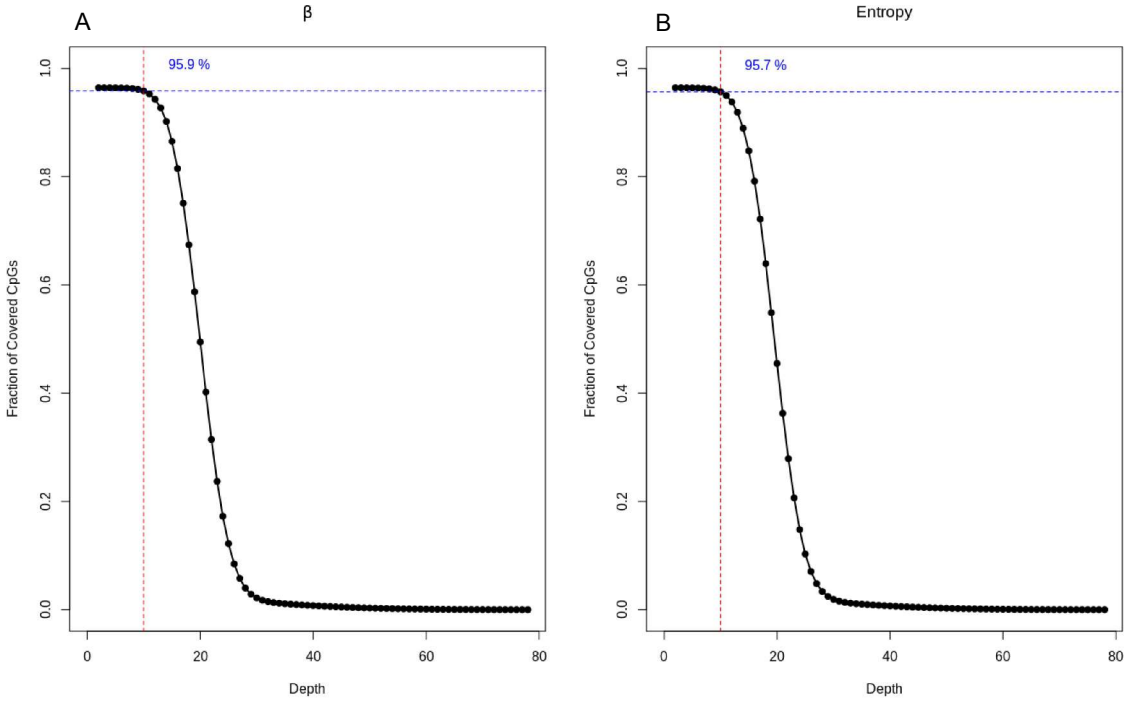

Figure S39: Figure shows an example of Quality Plot generated with PoreMeth2 function PoreMeth2PairedExpQualityPlot. The two quadrants show the distribution of the number of reads (depth) used to calculate β and S in each common CpG between the two experiments, with depth being the minimum value among the two samples for each CpG. The vertical and horizontal lines highlight the fraction of CpGs covered by at least 10 reads in both samples.