CpelTdm.jl: a Julia package for targeted differential DNA methylation analysis

Jordi Abante and John Goutsias

Whitaker Biomedical Engineering Institute and Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Motivation: Identifying regions of the genome that demonstrate significant differences in DNA methylation between groups of samples is an important problem in computational epigenetics. Available methods assume that methylation occurs in a statistically independent manner at individual cytosine-phosphate-guanine (CpG) sites or perform analysis using empirically estimated joint probability distributions of methylation patterns at no more than 4 contiguous CpG sites. These approaches can lead to poor detection performance and loss of reliability and reproducibility due to reduced specificity and sensitivity in the presence of insufficient data.

Results: To accommodate data obtained with different bisulfite sequencing technologies, such as RRBS, ERRBS, and WGBS, and improve statistical power, we developed CpelTdm.jl, a Julia package for targeted differential analysis of DNA methylation stochasticity between groups of unmatched or matched samples. This package performs rigorous statistical analysis of methylation patterns within regions of the genome specified by the user that takes into account correlations in methylation and results in robust detection of genomic regions exhibiting statistically significant differences in methylation stochasticity. CpelTdm.jl does not only detect mean methylation differences, as it is commonly done by previous methods, but also differences in methylation entropy and, more generally, between probability distributions of methylation.

Availability and Implementation: This Julia package is supported for Windows, MacOS, and Linux, and can be freely downloaded from GitHub: <https://github.com/jordiabante/CpelTdm.jl>.

Contacts: jabante1@jhu.edu or goutsias@jhu.edu.

Introduction

Differential analysis of DNA methylation using bisulfite sequencing data is an important task in epigenetic studies. Although many statistical approaches have been developed for this task (Landan *et al.*, 2012; Li *et al.*, 2014; Robinson *et al.*, 2014), most methods do not directly account for correlations in methylation, while others cannot reliably estimate joint probability distributions (PDMs) of methylation patterns, due to the lack of sufficient data to properly estimate a geometrically growing number of parameters, leading to reduced specificity and sensitivity as well as loss of reliability and reproducibility (Jenkinson *et al.*, 2017, 2018). To deal with these issues, Jenkinson *et al.* (2017) developed informME, a method for genome-wide analysis of methylation stochasticity that uses concepts from statistical physics and information theory. To achieve a computationally manageable approach to genome-wide methylation analysis, informME characterizes DNA methylation using the probabilities of methylation levels (percentage of methylated CpG sites) within non-overlapping genomic units composed of 150 bp each, and performs analysis at the resolution of one genomic unit. Consequently, this method does not allow direct comparisons between PDMs, which can provide valuable insights into mechanisms underlying the formation of methylation patterns (Landan *et al.*, 2012). In addition, informME employs a hypothesis testing approach based on a null distribution that is computed from control samples, thus introducing directionality in the analysis that is not always desirable. The genome-wide nature of methylation analysis performed by informME may also reduce its power to reject the null hypothesis when the alternative hypothesis is true, potentially leading to false negatives, a problem that can be remedied by performing targeted analysis on specific genomic regions of interest.

To address these issues, we developed a differential method for methylation pattern analysis by using a recently introduced correlated potential energy landscape (CPEL) model (Abante *et al.*, 2020) and by employing non-directional permutation methods for hypothesis testing. This approach, which is implemented in the freely available Julia package CpelTdm.jl, allows a user to perform targeted differential methylation (TDM) analysis of the stochastic behavior of methylation patterns between two groups of *unmatched* or *matched* bisulfite (RRBS, ERRBS, or WGBS) sequencing data in a statistically rigorous manner.

Approach

CpelTdm.jl involves several steps; see Fig. 1 and the Methods section. First, it uses a CPEL approach to characterize DNA methylation within analysis regions that are determined from a user-specified single-track input BED file. This approach models the stochastic methylation

state at all CpG sites within an analysis region via a PDM specified by a small number of parameters, which accounts for correlations among the methylation states at contiguous CpG sites. It subsequently uses independent and possibly partially observed bisulfite sequencing reads of the methylation state and estimates the values of the PDM parameters via maximum-likelihood. Finally, it uses the estimated CPEL models to perform differential analysis between two groups of unmatched or matched data samples by testing the null hypothesis that no differences in the information content of the methylation state are present in a given genomic region. This is done by defining appropriate differential test statistics in terms of the mean methylation level (MML), which evaluates the fraction of methylated CpG sites within an analysis region, the normalized methylation entropy (NME), which quantifies the amount of methylation stochasticity (pattern heterogeneity), and the coefficient of methylation divergence (CMD), which quantifies PDM differences using the information-theoretic notions of Kullback-Leibler divergence and cross entropy; see Figs. 2 & 3 and the Methods section.

The main challenge in performing hypothesis testing is computing the null distribution of a test statistic from given data when a theoretical distribution is not available. To address this problem and for computational efficiency, CpelTdm.jl performs permutation testing using one of two approaches depending on the number L of all possible data permutations: (exact) randomization testing is performed when $L < 1000$, and (approximate) Monte Carlo based permutation testing is performed when $L \geq 1000$; see the Methods section. Both approaches properly control the Type I error (false positives), in the sense that the probability of such an error is no larger than the test’s significance level. The resulting P -values are finally corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure for FDR control.

Results

Using simulations, we investigated the effectiveness of CpelTdm.jl for correctly identifying significant differences in methylation stochasticity between two groups containing 5 or 7 unmatched and 6 or 10 matched samples each. We considered a small genomic region with 4 CpG sites and characterized methylation stochasticity within the region by a CPEL model with parameters α and β of known ‘true’ values leading to the following four scenarios: no methylation differences, differences in MML and PDM, differences in NME and PDM, and differences in MML, NME, and PDM.

We generated methylation data in agreement with the number of samples associated with each group. Each data sample was associated with R methylation reads obtained by sampling the ‘true’ CPEL model corresponding to each group, where the value of R was randomly determined

to be between 10 and 30 with equal probability. Using these data, we computed an ‘estimated’ CPEL model for each sample, from which we calculated test statistic and corresponding P -values in a specific group comparison. Finally, by repeating this process 1000 times, we derived empirical estimates of the cumulative distribution functions of the P -values for each differential test statistic; see Figs. 4-7. The results showed the ability of CpelTdm.jl to detect statistical significance at a high rate (at least 95% of the time) while keeping the probability of false detections below 5% in all cases, thus demonstrating its reliability for correctly performing differential methylation analysis.

We also performed an unmatched-pair comparison over CpG islands (CGIs) between a group of 7 acute promyelocytic leukemia (APL) samples and a group of 6 remission samples after treatment using a previously published RRBS dataset; see Table 1. Densities of MML and NME values demonstrated an overall reduction of methylation level in the remission samples, consistently with previous results (Hebestreit *et al.*, 2013), as well as a reduction in methylation entropy leading to a more ordered methylation landscape after treatment; see Fig. 8. This behavior was also reflected by the distributions of differential MML and NME test statistic values and was associated with differences in PDM; see Fig. 9a. Q-Q plots obtained by differential analysis for each test statistic revealed that most observed P -values below 0.05 were much smaller than the corresponding expected P -values under the null hypothesis, providing strong evidence of the existence of methylation differences between the two groups; see Fig. 9b. Multiple hypothesis testing identified a number of CGIs exhibiting significant differences in MML, NME, and PDM (Q -value ≤ 0.1), which were linked to the promoter regions of 618 genes; see Fig. 10 and Table 2. Interestingly, gene set enrichment analysis (GSEA) of the list of genes exhibiting significant PDM differences at their promoter regions revealed enrichments related to development, differentiation, morphogenesis, signaling, transcription, and chromatin/chromosome formation and organization (see Table 3), pointing to their importance in APL.

The previous APL analysis took about 7 hours using 22 CPUs and about 1 GB of RAM per CPU. The required input CGI BED file and the output bedGraph files can be freely downloaded from http://www.cis.jhu.edu/~goutsias/data/CpelTdm_output.zip.

Methods

Defining analysis regions. CpelTdm.jl (see Fig. 1 for a flowchart) performs targeted differential DNA methylation analysis in order to identify statistically significant differences in methylation stochasticity between two groups of samples within genomic regions of interest that may include CpG islands (CGIs), promoter regions, enhancers, CpG clusters computed by the method

proposed by Hebestreit *et al.* (2013), etc. These regions must be specified by the user in a single-track input BED file (see Fig. 1). To strike a balance between computational cost and estimation performance, CpelTdm.jl partitions any connected target genomic region whose size is more than 1-kb into a minimum number of equally-sized subregions of size ≤ 1 -kb. It then independently performs downstream methylation analysis only within subregions for which the coverage at each CpG site is at least 5, which we refer to as ‘analysis regions’.

Modeling DNA methylation stochasticity. CpelTdm.jl models methylation stochasticity within an analysis region $\mathcal{R}$ that contains N CpG sites $n = 1, 2, \dots, N$ using the probability distribution of methylation (PDM) $\pi(\mathbf{x}) = \Pr[\mathbf{X} = \mathbf{x}]$ of an $N \times 1$ random vector $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_N]^T$, where $X_n = 0$ if the n -th CpG site is unmethylated and $X_n = 1$ if methylated. It then partitions the analysis region $\mathcal{R}$ into the minimum number of equally-sized non-overlapping subregions $\mathcal{R}_k$, $k = 1, 2, \dots, K$, such that the size of each subregion does not exceed a maximum size G (in bp) and characterizes the random methylation state $\mathbf{X}$ by means of a previously introduced correlated potential energy landscape (CPEL) model, given by (Abante *et al.*, 2020)

$$\pi(\mathbf{x}) = \frac{1}{Z} \exp\{-U(\mathbf{x})\}, \text{ for every } \mathbf{x} \in \{0, 1\}^N, \quad (1)$$

with potential energy function

$$U(\mathbf{x}) = - \sum_{k=1}^K \alpha_k \sum_{n \in \mathcal{N}_k} (2x_n - 1) - \beta \sum_{n=1}^{N-1} (2x_n - 1)(2x_{n+1} - 1), \quad (2)$$

and partition function

$$Z = \sum_{\mathbf{x}} \exp\{-U(\mathbf{x})\}, \quad (3)$$

where $\mathcal{N}_k$ denotes the set of all CpG sites within $\mathcal{R}_k$, α_k is a parameter that influences the propensity of CpG sites in the k -th subregion $\mathcal{R}_k$ to be methylated due to non-cooperative factors, and β is a parameter that accounts for the fact that the methylation status of two contiguous CpG sites n and $n + 1$ within the entire analysis region $\mathcal{R}$ could be correlated. Notably, parameter G determines the ‘granularity’ of modeling with its value chosen in a way that strikes a balance between the scale of differential methylation analysis and the success rate of estimating the parameters of the CPEL model from available bisulfite data. CpelTdm.jl uses a default granularity of $G = 250$ bp and performs necessary partition function computations and PDM marginalizations using a set of previously developed computationally efficient equations. These equations involve multiplications of 2×2 matrices determined by spectral decompositions requiring computation of eigenvalues and eigenvectors, which are performed analytically using exact formulas (Abante *et al.*, 2020).

Performing parameter estimation. Given I independent and possibly partially observed reads $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I$ of the methylation state in a given analysis region, CpelTdm.jl estimates the parameters $\boldsymbol{\theta} = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_K \ \beta]^T$ of the potential energy function of the CPEL model (see Fig. 1) by solving the following maximum-likelihood optimization problem:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^I \ln \pi(\mathbf{x}_i; \boldsymbol{\theta}). \quad (4)$$

This is done by using Simulated Annealing with a temperature reduction factor of 10^{-4} , a global optimization algorithm that was previously determined to yield superior estimation performance when compared to a number of alternative methods (Abante *et al.*, 2020).

Computing mean methylation levels. CpelTdm.jl quantifies the average amount of methylation within an analysis region containing N CpG sites using the mean methylation level (MML), given by (Jenkinson *et al.*, 2017)

$$\mu(\mathbf{X}) = \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N X_n \right], \quad (5)$$

where $\mathbb{E}[\cdot]$ denotes expectation. The MML evaluates the fraction of CpG sites that are methylated within an analysis region, taking its minimum value when all CpG sites are unmethylated and achieving its maximum value when all CpG sites are methylated; see Fig. 2. It can be shown that (Abante *et al.*, 2020)

$$\mu(\mathbf{X}) = \frac{1}{2} \left(1 + \frac{1}{N} \sum_{k=1}^K \frac{\partial \ln Z}{\partial \alpha_k} \right), \quad (6)$$

a formula employed by CpelTdm.jl to efficiently evaluate the MML by computing the derivatives of the logarithm of the partition function with respect to the parameters α_k using a standard derivative approximation technique.

Computing normalized methylation entropies. CpelTdm.jl quantifies the amount of methylation stochasticity (pattern heterogeneity) within an analysis region containing N CpG sites using the normalized methylation entropy (NME), given by (Jenkinson *et al.*, 2017)

$$h(\mathbf{X}) = -\frac{1}{N} \sum_{\mathbf{x}} \pi(\mathbf{x}) \log_2 \pi(\mathbf{x}). \quad (7)$$

The NME ranges between 0 and 1, taking its minimum value when a single methylation pattern is observed within the analysis region (perfectly ordered methylation) and achieving its maximum value when all methylation patterns are equally likely (fully disordered methylation); see Fig. 2. It can be shown that (Abante *et al.*, 2020)

$$h(\mathbf{X}) = \frac{1}{N \ln 2} \left(\ln Z - \sum_{k=1}^K \alpha_k \frac{\partial \ln Z}{\partial \alpha_k} - \beta \frac{\partial \ln Z}{\partial \beta} \right), \quad (8)$$

a formula used by CpelTdm.jl to efficiently evaluate the NME by employing the previously computed derivatives of the logarithm of the partition function with respect to the parameters α_k and by also computing the derivative of the logarithm of the partition function with respect to parameter β using a standard derivative approximation technique.

Computing coefficients of methylation divergence. Within an analysis region, CpelTdm.jl quantifies differences between the PDMs $\pi_1(\mathbf{x})$ and $\pi_2(\mathbf{x})$ of the methylation states $\mathbf{X}_1$ and $\mathbf{X}_2$ in a data sample from the first group and a data sample from the second group by using the coefficient of methylation divergence (CMD), which we define by

$$Q(\mathbf{X}_1, \mathbf{X}_2) = \frac{D(\pi_1 \parallel \bar{\pi}) + D(\pi_2 \parallel \bar{\pi})}{C(\mathbf{X}_1, \bar{\mathbf{X}}) + C(\mathbf{X}_2, \bar{\mathbf{X}})}, \quad (9)$$

where $\bar{\mathbf{X}}$ is a random vector characterized by a CPEL model $\bar{\pi}$ whose potential energy function is the average of the potential energy functions associated with $\mathbf{X}_1$ and $\mathbf{X}_2$,

$$D(f \parallel g) = \sum_{\mathbf{x}} f(\mathbf{x}) \log_2 \frac{f(\mathbf{x})}{g(\mathbf{x})} \quad (10)$$

is the Kullback-Leibler (KL) divergence between two probability distributions f and g , and

$$C(\mathbf{X}, \mathbf{Y}) = - \sum_{\mathbf{x}} f(\mathbf{x}) \log_2 g(\mathbf{x}) \quad (11)$$

is the cross entropy between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ characterized by probability distributions f and g , respectively. Notably, $[D(\pi_1 \parallel \bar{\pi}) + D(\pi_2 \parallel \bar{\pi})]/2$ is known as the *geometric* Jensen-Shannon divergence (Nielsen, 2019), since it quantifies differences between the probability distributions π_1 and π_2 and their geometric average $\bar{\pi} \sim \sqrt{\pi_1 \pi_2}$, as compared to the standard Jensen-Shannon distance that quantifies differences between π_1 and π_2 and their arithmetic average $(\pi_1 + \pi_2)/2$. The CMD ranges between 0 and 1, taking its minimum value when the two PDMs π_1 and π_2 are identical and achieving its maximum when the supports of the probability distributions π_1 and π_2 do not overlap with the support of $\bar{\pi}$, indicating radically different PDMs between the two groups.

It can be shown that

$$Q(\mathbf{X}_1, \mathbf{X}_2) = 1 - \frac{h(\mathbf{X}_1) + h(\mathbf{X}_2)}{c(\mathbf{X}_1, \bar{\mathbf{X}}) + c(\mathbf{X}_2, \bar{\mathbf{X}})}, \quad (12)$$

where h is the NME and c is the normalized cross methylation entropy (i.e., the cross entropy divided by N). Moreover,

$$h(\mathbf{X}_1) = \frac{1}{N \ln 2} \left(\ln Z_1 - \sum_{k=1}^K \alpha_{1,k} \frac{\partial \ln Z_1}{\partial \alpha_{1,k}} - \beta_1 \frac{\partial \ln Z_1}{\partial \beta_1} \right), \quad (13)$$

and

$$c(\mathbf{X}_1, \mathbf{X}_2) = \frac{1}{N \ln 2} \left(\ln Z_2 - \sum_{k=1}^K \alpha_{2,k} \frac{\partial \ln Z_1}{\partial \alpha_{1,k}} - \beta_2 \frac{\partial \ln Z_1}{\partial \beta_1} \right), \quad (14)$$

where $\{\alpha_{1,k}, k = 1, 2, \dots, K, \beta_1, Z_1\}$ and $\{\alpha_{2,k}, k = 1, 2, \dots, K, \beta_2, Z_2\}$ are respectively the energy parameters and partition functions of the CPEL models associated with $\mathbf{X}_1$ and $\mathbf{X}_2$. Using these formulas, CpelTdm.jl efficiently evaluates the CMD by computing partition function derivatives using a standard approximation technique.

Performing hypothesis testing. CpelTdm.jl identifies regions of the genome that demonstrate statistically significant differences in DNA methylation stochasticity between two groups of samples by performing an unmatched or matched pairs group comparison. This is achieved by using permutation methods (Ernst, 2004) to test, for each analysis region, the null hypothesis H_0 that each pair of samples exhibits no difference in methylation stochasticity (quantified by the MML, NME, or PDM) regardless of the specific group sample assignment.

Unmatched pairs group comparison. In an unmatched pairs comparison associated with a group of M_1 values of a given statistical summary (MML, NME, or PDM) of methylation stochasticity within an analysis region and another group of M_2 values associated with the same analysis region, CpelTdm.jl performs hypothesis testing using the following differential test statistics (see Fig. 1):

$$T_{\text{MML}} = \frac{1}{M_1} \sum_{m=1}^{M_1} \mu(\mathbf{X}_1^{(m)}) - \frac{1}{M_2} \sum_{m=1}^{M_2} \mu(\mathbf{X}_2^{(m)}), \quad (15)$$

$$T_{\text{NME}} = \frac{1}{M_1} \sum_{m=1}^{M_1} h(\mathbf{X}_1^{(m)}) - \frac{1}{M_2} \sum_{m=1}^{M_2} h(\mathbf{X}_2^{(m)}), \quad (16)$$

and

$$T_{\text{PDM}} = \frac{1}{M_1 M_2} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} Q(\mathbf{X}_1^{(m_1)}, \mathbf{X}_2^{(m_2)}), \quad (17)$$

where $\mathbf{X}_1^{(m)}$ and $\mathbf{X}_2^{(m)}$ are the m -th methylation states of the analysis region within the first and second group, respectively. The test statistic T_{MML} quantifies the difference between the average of the mean methylation levels in each groups, T_{NME} assesses the difference between the averages of the normalized methylation entropies, and T_{PDM} quantifies the average of the PDM differences observed between the two groups; see Fig. 3.

For each test statistic T , a (two-tailed) hypothesis test requires knowledge of the null cumulative distribution function $F_0(t) = \Pr[|T| < t \mid H_0]$, which can then be used to calculate the P -value associated with an observation t^* of T by $p = 1 - F_0(|t^*|)$. To compute $F_0(t)$ for an analysis region, CpelTdm.jl uses a ‘randomization model’ that randomly assigns the available

samples to the two groups, thus forming $L = (M_1 + M_2)!/M_1!M_2!$ group assignments, provided that L is less than a maximum allowable number L_0 , which is taken to be $L_0 = 1000$ by default; see Fig. 1. Under the null hypothesis, the group assignments are equally likely (with probability $1/L$) to produce the expected outcome. As a consequence, the null cumulative distribution function $F_0(t)$ of a test statistic T will be given by $\frac{1}{L} \sum_{l=1}^L I[|t_l| < t]$, where t_l , $l = 1, 2, \dots, L$, are values of T computed for each group assignment and $I[\cdot]$ is the Iverson bracket, taking value 1 when its argument is true and 0 otherwise. This leads to an *exact* P -value computation given by (see Fig. 1)

$$p = 1 - \frac{1}{L} \sum_{l=1}^L I[|t_l| < |t^*|] = \frac{1}{L} \sum_{l=1}^L I[|t_l| \geq |t^*|], \quad (18)$$

where t^* is the value of the T statistic calculated from the given data.

Note that the only possible values for p produced by the previous randomization testing approach are $1/L, 2/L, \dots, 1$. Moreover, if the significance level of the test is set to $a = k/L$ for some integer k such that p can take value k/L , then

$$\begin{aligned} \Pr[\text{Type I error}] &= \Pr[p \leq a \mid H_0] = \Pr\left[\frac{1}{L} \sum_{l=1}^L I[|t_l| \geq |t^*|] \leq \frac{k}{L} \mid H_0\right] \\ &= \Pr\left[\sum_{l=1}^L I[|t_l| \geq |t^*|] \leq k \mid H_0\right] = \frac{k}{L} = a. \end{aligned} \quad (19)$$

On the other hand, if $a = k/L$ for some integer k such that p cannot take a value k/L , then $\Pr[\text{Type I error}] < a$ and the test will be conservative. As a consequence, the randomization testing procedure used by CpelTdm.jl can always control the Type I error (false positives).

When $L \geq L_0$, the previous method becomes computationally intensive. For this reason, CpelTdm.jl switches to a hypothesis testing approach that estimates the P -value using a permutation test based on Monte Carlo sampling; see Fig. 1. In this case, L distinct sample permutations are performed by assigning M_1 samples to the first group and the remaining samples to the second group. The null cumulative distribution function is then approximated by independently sampling, $L_0 - 1$ times, the set of L permutations with equal probability and by setting $\hat{F}_0(t) = \frac{1}{L_0} (I[|t^*| < t] + \sum_{l=1}^{L_0-1} I[|t_l| < t])$, since this method produces L_0 test statistic values, including the value t^* computed from the data. In this case, the P -value is approximated by (see Fig. 1)

$$\hat{p} = 1 - \frac{1}{L_0} \sum_{l=1}^{L_0-1} I[|t_l| < |t^*|] = \frac{1}{L_0} \left(1 + \sum_{l=1}^{L_0-1} I[|t_l| \geq |t^*|]\right). \quad (20)$$

Note that the only possible values for $\hat{p}$ are $1/L_0, 2/L_0, \dots, 1$, which implies that $\hat{p} \geq 0.001$ when $L_0 = 1000$. Moreover, if the significance level of the test is taken to be $a = k/L_0$ for

some integer k such that $\hat{p}$ can take value k/L_0 , then it can be shown that $\Pr[\text{Type I error}] = a$, whereas if the significance level is taken to be $a = k/L_0$ for some integer k such that $\hat{p}$ cannot take a value k/L_0 , then $\Pr[\text{Type I error}] < a$. Consequently, this procedure can also control the Type I error.

Matched pairs group comparison. Assuming availability of M pairs of matched values of a given statistical summary from the first and second groups, associated with an analysis region, CpelTdm.jl performs hypothesis testing using the previous randomization testing method, provided that the total number $L = 2^M$ of possible matched group assignments is less than L_0 . In this case, the M matched pairs are used to form L distinct permutations, each containing all M pairs but with some group labels being reversed, and a value t_l is computed for the l -th permutation using each of the following two differential test statistics:

$$T_{\text{MML}} = \frac{1}{M} \sum_{m=1}^M \left[\mu(\mathbf{X}_1^{(m)}) - \mu(\mathbf{X}_2^{(m)}) \right], \quad (21)$$

and

$$T_{\text{NME}} = \frac{1}{M} \sum_{m=1}^M \left[h(\mathbf{X}_1^{(m)}) - h(\mathbf{X}_2^{(m)}) \right]. \quad (22)$$

However, if $L \geq L_0$, CpelTdm.jl performs hypothesis testing using the Monte Carlo based permutation procedure employed in the unmatched case in which the $L_0 - 1$ values of the test statistic required by the method are determined by independently sampling the set of all L matched group permutations with equal probability and by computing the test statistic value for each permutation.

Unfortunately, the previous procedure cannot be used to perform hypothesis testing using the CMD because of its symmetry; i.e., due to the fact that $Q(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)}) = Q(\mathbf{X}_2^{(m)}, \mathbf{X}_1^{(m)})$. For this reason, CpelTdm.jl simply calculates the average of the CMDs associated with all matched analysis regions, given by $1/M \sum_{m=1}^M Q(\mathbf{X}_1^{(m)}, \mathbf{X}_2^{(m)})$, and outputs the results.

Note that, in both the unmatched and matched cases, CpelTdm.jl performs hypothesis testing only when the number of group permutations L is large enough ($L \geq 20$) to produce P -values that are not above a 5% significance level (recall that, when $L < 1000$, the P -values cannot be less than $1/L$). Finally, CpelTdm.jl performs multiple hypothesis testing using Benjamini-Hochberg (BH) correction for false discovery rate (FDR) control (see Fig. 1) and provides the adjusted p -values in the output.

Funding

This work was supported by NSF Grant EFRI CEE 132452 and NIH-NHGRI Grant RM1HG008529. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Abante, J. *et al.* (2020) Detection of haplotype-dependent allele-specific DNA methylation in WGBS data, *Nat. Commun.*, **11**, 5238.
- Robinson, M. D. *et al.* (2014) Statistical methods for detecting differentially methylated loci and regions, *Front. Genet.*, **5**, 324.
- Hebestreit, K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data, *Bioinformatics*, **29**, 1647-1653.
- Landan, G. *et al.* (2012) Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues, *Nat. Genet.*, **44**, 1207-1214.
- Li, S. *et al.* (2014) Dynamic evolution of clonal epialleles revealed by methclone, *Genome Biol.*, **15**, 472.
- Jenkinson, G. *et al.* (2017) Potential energy landscapes identify the information-theoretic nature of the epigenome, *Nat. Genet.*, **49**, 719-729.
- Jenkinson, G. *et al.* (2018) An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data, *BMC Bioinf.*, **19**, 87.
- Hebestreit, K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data, *Bioinformatics*, **29**, 1647-1653.
- Nielsen, F. (2019) On the Jensen-Shannon symmetrization of distances relying on abstract means, *Entropy*, **21**, 485.
- Ernst, M. D. (2004) Permutation methods: A basis for exact inference, *Stat. Sci.*, **19**, 676-685.

Figure 1. Flowchart of CpelTdm.jl for differential methylation analysis between two groups of unmatched samples in terms of the mean methylation level (MML) μ , the normalized methylation entropy (NME) h , and the coefficient of methylation divergence (CMD) Q . A similar approach is employed in the case of matched samples.

Figure 2. Heat maps of the mean methylation level (MML) and the normalized methylation entropy (NME) within an analysis region containing 4 CpG sites whose methylation state is characterized by a CPEL model with two parameters $-3 \leq \alpha \leq 3$ and $0 \leq \beta \leq 3$. Smaller negative values of α result in decreased MML values, indicating that fewer CpG sites are methylated on the average, whereas larger positive values result in increased MML values, indicating that more CpG sites are methylated on the average. This behavior is associated with increasingly reduced NME values in both cases, and is exacerbated with larger values of β due to increased correlation. When $\alpha = \beta = 0$, each CpG site is methylated independently from each other with 50% probability, resulting in an equal number of CpG sites to be methylated and unmethylated on the average (MML = 0.5) and maximum methylation entropy (NME = 1).

Figure 3. Examples of differential test statistic values t_{MML} , t_{NME} , and t_{PDM} between a reference and a test analysis region containing 4 CpG sites whose methylation state is characterized by a CPEL model with two parameters α and β . (a) Reference model: $\alpha = -4$, $\beta = 0$; test models: $-4 \leq \alpha \leq 4$, $\beta = 0$. (b) Reference model: $\alpha = -4$, $\beta = 0.5$; test models: $-4 \leq \alpha \leq 4$, $\beta = 0.5$. (c) Reference model: $\alpha = -0.5$, $\beta = 4$; test models: $\alpha = -0.5$, $0 \leq \beta \leq 4$. (d) Reference model: $\alpha = 0$, $\beta = 4$; test models: $\alpha = 0$, $0 \leq \beta \leq 4$. The CpG sites in the reference analysis region are unmethylated in (a), (b), and (c), whereas half of them are methylated in (d) on the average. Notably, the test region transitions from being unmethylated to being methylated as α increases in (a) and (b), and to being partially methylated as β decreases in (c). Interestingly, no MML difference is present in (d), despite differences in NME and PDM. Note also that methylation is uncorrelated in (a), and this is also true in (c) and (d) when $\beta = 0$. On the average, most CpG sites in the test analysis region are methylated ($t_{MML} > 0.5$) in (a) and (b) when $\alpha > 0$ and unmethylated ($t_{MML} < 0.5$) when $\alpha < 0$. Moreover, the differential test statistic T_{NME} achieves its maximum value when $\alpha = \beta = 0$ in (a), corresponding to a CPEL model which assumes that the CpG sites in the test analysis region are independently methylated with 50% probability.

Figure 4. Monte Carlo estimation of the cumulative distribution functions of the P -values obtained by CpelTdm.jl when performing *randomization testing* using the differential test statistics T_{MML} , T_{NME} , and T_{PDM} . A simulated *unmatched pairs group comparison* is considered within a small genomic region containing 4 CpG sites when the test and reference group contain 5 data samples each ($L = 252 < 1000$). (a) Region exhibits no differences (test: $\alpha = 0.5, \beta = 0.5$; reference: $\alpha = 0.5, \beta = 0.5$). (b) Region exhibits MML and PDM differences (test: $\alpha = 0.5, \beta = 0.5$; reference: $\alpha = -0.5, \beta = 0.5$). (c) Region exhibits NME and PDM differences (test: $\alpha = 0.0, \beta = 0.5$; reference: $\alpha = 0, \beta = 0$). (d) Region exhibits MML, NME, and PDM differences (test: $\alpha = 0.5, \beta = 0.5$; reference: $\alpha = 0, \beta = 0$). The marked (red stars) cumulative distribution functions indicate that CpelTdm.jl detects statistically significant differences with a rate that is no less than 95% (i.e., $\Pr[p \leq 0.05] \geq 0.95$), in agreement with the region's known differential behavior. In the remaining cases, $\Pr[p \leq 0.05] = 0.05$, indicating a 5% probability of Type I error.

Figure 5. Monte Carlo estimation of the cumulative distribution functions of the P -values obtained by CpelTdm.jl when performing *Monte Carlo based permutation testing* using the differential test statistics T_{MML} , T_{NME} , and T_{PDM} . A simulated *unmatched pairs group comparison* is considered within a small genomic region containing 4 CpG sites when the test and reference group contain 7 data samples each ($L = 3432 > 1000$). (a) Region exhibits no differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0.5$, $\beta = 0.5$). (b) Region exhibits MML and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = -0.5$, $\beta = 0.5$). (c) Region exhibits NME and PDM differences (test: $\alpha = 0.0$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). (d) Region exhibits MML, NME, and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). The marked (red stars) cumulative distribution functions indicate that CpelTdm.jl detects statistically significant differences 100% of the time (i.e., $\Pr[p \leq 0.05] = 1$), in agreement with the region's known differential behavior. In the remaining cases, $\Pr[p \leq 0.05] = 0.05$, indicating a 5% probability of Type I error.

Figure 6. Monte Carlo estimation of the cumulative distribution functions of the P -values obtained by CpelTdm.jl when performing *randomization testing* using the differential test statistics T_{MML} and T_{NME} . A simulated *matched pairs group comparison* is considered within a small genomic region containing 4 CpG sites when the test and reference group contain 6 data samples each ($L = 64 < 1000$). (a) Region exhibits no differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0.5$, $\beta = 0.5$). (b) Region exhibits MML and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = -0.5$, $\beta = 0.5$). (c) Region exhibits NME and PDM differences (test: $\alpha = 0.0$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). (d) Region exhibits MML, NME, and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). The marked (red stars) cumulative distribution functions indicate that CpelTdm.jl detects statistically significant differences with a rate that is no less than 95% (i.e., $\Pr[p \leq 0.05] \geq 0.95$), in agreement with the region's known differential behavior. In the remaining cases, $\Pr[p \leq 0.05] = 0.05$, indicating a 5% probability of Type I error.

Figure 7. Monte Carlo estimation of the cumulative distribution functions of the P -values obtained by CpelTdm.jl when performing *Monte Carlo based permutation testing* using the differential test statistics T_{MML} and T_{NME} . A simulated *matched pairs group comparison* is considered within a small genomic region containing 4 CpG sites when the test and reference group contain 10 data samples each ($L = 1024 > 1000$). (a) Region exhibits no differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0.5$, $\beta = 0.5$). (b) Region exhibits MML and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = -0.5$, $\beta = 0.5$). (c) Region exhibits NME and PDM differences (test: $\alpha = 0.0$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). (d) Region exhibits MML, NME, and PDM differences (test: $\alpha = 0.5$, $\beta = 0.5$; reference: $\alpha = 0$, $\beta = 0$). The marked (red stars) cumulative distribution functions indicate that CpelTdm.jl detects statistically significant differences 100% of the time (i.e., $\Pr[p \leq 0.05] = 1$), in agreement with the region's known differential behavior. In the remaining cases, $\Pr[p \leq 0.05] = 0.05$, indicating a 5% probability of Type I error.

Figure 8. Densities of mean methylation level (MML) and normalized methylation entropy (NME) values within CGI analysis regions computed by CpelTdm.jl in an unmatched pairs group comparison between acute promyelocytic leukemia (APL) and remission samples. Notably, treatment results in an overall reduction of methylation level and methylation entropy leading to a more ordered methylation landscape after treatment.

Figure 9. (a) Distributions of the differential MML, NME, and PDM test statistic values t_{MML} , t_{NME} , and t_{PDM} obtained by CpelTdm.jl from 5398 CGI analysis regions in the unmatched pairs group comparison between acute promyelocytic leukemia and remission samples. (b) Q-Q plots for each differential test statistic (green lines) of quantiles of $-\log_{10} P$ -values observed within the CGI analysis regions versus quantiles of expected $-\log_{10} P$ -values under the null hypothesis. Notably, since most observed P -values below 0.05 (over the dotted line) are located above the diagonal, they are much smaller than the corresponding P -values expected under the null hypothesis, thus providing strong evidence of the existence of meaningful differences in methylation stochasticity among the two groups.

Figure 10. (a) Multiple hypothesis testing performed by CpelTdm.jl in the unmatched pairs group comparison between acute promyelocytic leukemia and remission samples identified CGI analysis regions with significant MML, NME, and PDM differences (Q -values ≤ 0.1). Notably, an unmatched pairs remission/remission group comparison did not find statistically significant analysis regions, as expected. (b) A total of 618 genes were identified to exhibit statistically significant MML, NME, or PDM differences within at least one CGI analysis region overlapping their promoter regions (4-kb window centered at TSS). Notably, 495 genes (80%) were associated with significant MML differences (dMML genes), 399 (65%) with significant NME differences (dNME genes), and 574 (93%) with significant PDM differences (dPDM genes). In addition, 22 genes were associated with only significant MML differences (e.g., the homeobox gene *HOXD8* implicated in stem cell differentiation and cancer, including APL, the proto-oncogene *RET* involved in activating signaling pathways that play a role in cell differentiation, growth, migration and survival, and the transcription factor *TCF7L1* implicated in gastric and colorectal cancers as well as in acute lymphoblastic leukemia). Moreover, 8 genes were associated with only significant NME differences (e.g., *IGFBP5* that is often dysregulated in cancer, *LAPTM4B*, an oncogene that has been found to be highly expressed in various tumors, and *PAWR*, a tumor suppressor that induces apoptosis in cancer cells). Finally, 109 genes were associated with only significant PDM differences (e.g., *AK1* and *SYNE2*, both implicated in acute myeloid leukemia, *ITPKA*, which is expressed in a broad range of tumor types but shows limited expression in normal cells, and *MPPED2* that plays a carcinogenetic role across different cancer types).

Table 1. Summary of the RRBS dataset used for targeted differential methylation analysis, which includes 7 acute promyelocytic leukemia (APL) samples and 6 remission samples (NCBI GEO accession number GSE42044).

Name	Sample	Condition	Sex
apl1	APL_10759	APL	Female
apl2	APL_7717	APL	Female
apl3	APL_10545	APL	Male
apl4	APL_19	APL	Male
apl5	APL_9882	APL	Male
apl6	APL_5	APL	Female
apl7	APL_10892	APL	Female
remission1	APL_10961	Remission	Female
remission2	APL_11624	Remission	Female
remission3	APL_11523	Remission	Male
remission4	APL_16	Remission	Male
remission5	APL_10325	Remission	Male
remission6	APL_5894	Remission	Male

Table 2. Genes identified by CpelTdm.jl to exhibit significant MML, NME, or PDM differences (dMML, dNME, and dPDM genes) within at least one CGI analysis region overlapping their promoter regions (4-kb window centered at TSS) in the unmatched pairs group comparison between the acute promyelocytic leukemia and remission samples.

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>ABCA17P</i>			*	<i>ARFGEF3</i>	*		*
<i>ABCA3</i>	*	*	*	<i>ARHGAP27P2</i>			*
<i>ABCC8</i>	*	*	*	<i>ARHGAP39</i>	*	*	*
<i>ABO</i>	*		*	<i>ARHGEF10</i>			*
<i>ABTB1</i>	*		*	<i>ARHGEF12</i>	*	*	*
<i>ACOT12</i>	*	*		<i>ARHGEF26</i>	*	*	*
<i>ACTL6B</i>	*	*	*	<i>ARNT2</i>	*	*	*
<i>ACTN3</i>	*		*	<i>ASB3</i>	*	*	*
<i>ACTR3C</i>	*	*	*	<i>ASCL5</i>	*		*
<i>ADAM23</i>	*		*	<i>ATP6V0E2-AS1</i>	*	*	*
<i>ADAM32</i>	*	*	*	<i>BAALC</i>			*
<i>ADAM8</i>	*	*	*	<i>BAX1</i>	*	*	*
<i>ADAMTSL3</i>			*	<i>BCKDK</i>		*	*
<i>ADAMTSL4</i>	*	*	*	<i>BCL11B</i>			*
<i>ADAP2</i>			*	<i>BEND4</i>			*
<i>ADCY2</i>	*	*	*	<i>BHLHE22</i>	*	*	*
<i>ADCY5</i>	*	*	*	<i>BICDL2</i>	*		*
<i>ADCYAP1R1</i>	*	*	*	<i>BMP2</i>	*		*
<i>ADGRB1</i>	*	*	*	<i>BMP7</i>	*	*	*
<i>ADORA2A-AS1</i>	*	*	*	<i>BOK</i>	*	*	*
<i>AFAP1L2</i>	*	*	*	<i>BRSK2</i>	*		*
<i>AGAP1</i>	*	*	*	<i>BX</i>	*	*	*
<i>AGAP3</i>			*	<i>BTG3</i>	*		*
<i>AHNAK2</i>	*			<i>C12orf42</i>	*	*	*
<i>AJAP1</i>	*	*	*	<i>C16orf91</i>	*	*	*
<i>AJUBA</i>	*	*	*	<i>C17orf102</i>	*	*	*
<i>AK1</i>			*	<i>C17orf58</i>			*
<i>ALDH1A3</i>	*		*	<i>C17orf82</i>			*
<i>ALX12</i>	*		*	<i>C1QL1</i>	*	*	*
<i>ALPK3</i>	*		*	<i>C1QTNF12</i>	*	*	*
<i>AX3</i>	*	*	*	<i>C2orf72</i>	*	*	
<i>AMER3</i>	*	*	*	<i>C5orf49</i>			*
<i>AMH</i>	*	*	*	<i>CA4</i>	*	*	*
<i>ANKRD65</i>	*	*	*	<i>CACHD1</i>	*	*	*
<i>ANO1</i>	*	*	*	<i>CACNA1A</i>	*	*	*
<i>ANO5</i>	*	*	*	<i>CACNB2</i>	*	*	*
<i>APBA2</i>	*	*	*	<i>CADPS2</i>	*		

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>CALCR</i>	*	*	*	<i>CPEB1</i>	*	*	*
<i>CAMSAP1</i>	*	*	*	<i>CPLX2</i>			*
<i>CBLC</i>	*	*	*	<i>CPNE7</i>	*	*	*
<i>CBLN1</i>	*	*	*	<i>CPNE9</i>			*
<i>CBLN4</i>	*	*	*	<i>CRHBP</i>	*	*	*
<i>CBS</i>			*	<i>CRLF1</i>	*		*
<i>CCBE1</i>	*	*	*	<i>CRMP1</i>	*	*	*
<i>CCDC166</i>	*	*	*	<i>CRTAC1</i>	*	*	*
<i>CCDC170</i>	*	*	*	<i>CTNNA1</i>	*		
<i>CCDC180</i>	*	*	*	<i>CTNND2</i>	*		*
<i>CCDC92</i>			*	<i>CWH43</i>	*		*
<i>CDC42EP5</i>	*	*	*	<i>CYB561</i>			*
<i>CDCP1</i>	*	*	*	<i>CYGB</i>	*	*	*
<i>CDH3</i>	*	*	*	<i>DAAM2</i>	*	*	*
<i>CDH4</i>	*		*	<i>DBI</i>			*
<i>CDX1</i>	*		*	<i>DBX2</i>	*	*	*
<i>CDX2</i>	*	*	*	<i>DEGS2</i>	*		*
<i>CEBPA-DT</i>	*	*	*	<i>DENND2A</i>			*
<i>CEND1</i>	*	*	*	<i>DHH</i>	*	*	*
<i>CHP2</i>	*	*	*	<i>DIO3</i>	*		*
<i>CHRD12</i>	*	*	*	<i>DIPK1B</i>	*	*	
<i>CHST10</i>			*	<i>DLL4</i>			*
<i>CHST3</i>			*	<i>DNAH10</i>			*
<i>CHST8</i>	*	*	*	<i>DNAJA4</i>	*		*
<i>CLDN23</i>			*	<i>DNER</i>	*		*
<i>CLIC6</i>	*	*	*	<i>DOCK1</i>	*	*	*
<i>CMIP</i>	*	*		<i>DOCK9</i>	*	*	*
<i>CMTM6</i>			*	<i>DPP10-AS1</i>			*
<i>CNGA3</i>	*		*	<i>DPYSL5</i>	*	*	*
<i>CNNM3</i>			*	<i>DRAXIN</i>	*	*	*
<i>CNTN4</i>	*	*	*	<i>DSCAM</i>	*	*	*
<i>COBL</i>		*	*	<i>DSP</i>			*
<i>COL13A1</i>	*	*	*	<i>EBF4</i>			*
<i>COL2A1</i>	*	*	*	<i>ECE1</i>	*	*	*
<i>COL4A2</i>	*	*	*	<i>EFNA5</i>	*	*	*
<i>COL4A4</i>	*	*	*	<i>EFNB2</i>	*	*	*
<i>COL9A3</i>	*	*	*	<i>EFS</i>	*		*
<i>COMP</i>	*	*	*	<i>ELOVL2</i>	*		
<i>COMTD1</i>	*	*	*	<i>ELOVL7</i>			*
<i>COQ7</i>			*	<i>EMILIN3</i>	*	*	*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>EML1</i>	*	*	*	<i>FREM2</i>	*	*	*
<i>EMX1</i>	*	*	*	<i>FSTL1</i>	*	*	*
<i>EN2</i>	*	*	*	<i>FSTL4</i>	*	*	*
<i>ENTPD2</i>	*	*	*	<i>FZD10-AS1</i>	*	*	*
<i>EPB41L3</i>	*	*	*	<i>FZD6</i>	*	*	*
<i>EPB41L4B</i>			*	<i>GABBR2</i>	*	*	*
<i>EPHA8</i>	*	*	*	<i>GABRA5</i>	*	*	*
<i>EPHX3</i>	*	*	*	<i>GABRG3</i>	*		*
<i>EPO</i>	*	*	*	<i>GAD1</i>	*	*	*
<i>EPS8L2</i>	*	*	*	<i>GALNT16</i>	*		*
<i>ERC2</i>	*	*	*	<i>GALNT17</i>	*	*	*
<i>ESRP1</i>	*	*	*	<i>GATA5</i>	*	*	*
<i>EVC2</i>	*	*	*	<i>GBX2</i>	*	*	*
<i>EVX1</i>	*	*	*	<i>GDF6</i>	*		*
<i>EYA4</i>	*	*	*	<i>GDNF</i>	*	*	*
<i>F11R</i>	*		*	<i>GFRA2</i>	*	*	*
<i>F2RL1</i>	*	*	*	<i>GGN</i>	*		*
<i>F3</i>	*			<i>GJD2</i>	*	*	*
<i>FAM13C</i>	*	*	*	<i>GLB1L2</i>	*	*	*
<i>FAM173A</i>	*	*		<i>GMDS</i>			*
<i>FAM20A</i>			*	<i>GMNC</i>	*	*	*
<i>FAM222A</i>			*	<i>GMPPA</i>	*		*
<i>FAM78A</i>	*	*		<i>GMPS</i>	*	*	*
<i>FAM83F</i>	*	*	*	<i>GNAS-AS1</i>		*	*
<i>FAM83H</i>	*	*	*	<i>GNMT</i>	*		*
<i>FANK1</i>			*	<i>GOLGA7B</i>			*
<i>FBXL16</i>	*	*	*	<i>GPR135</i>			*
<i>FENDRR</i>	*		*	<i>GPR158</i>	*	*	*
<i>FGF19</i>	*	*	*	<i>GPR37</i>	*		*
<i>FGF20</i>	*	*	*	<i>GPRIN1</i>	*	*	*
<i>FGF4</i>	*	*	*	<i>GPRIN2</i>	*	*	*
<i>FHOD3</i>	*	*	*	<i>GRB14</i>	*	*	*
<i>FMN2</i>	*	*	*	<i>GREB1L</i>	*	*	*
<i>FMNL3</i>			*	<i>GREM1</i>	*	*	*
<i>FNDC1</i>	*			<i>GRIA2</i>	*	*	*
<i>FNDC10</i>	*		*	<i>GRIK3</i>	*		*
<i>FOXC2</i>	*	*	*	<i>GRIK4</i>	*	*	*
<i>FOXE1</i>	*	*	*	<i>GRIN3B</i>			*
<i>FOXH1</i>	*	*	*	<i>GRM1</i>	*	*	*
<i>FOXL1</i>	*	*	*	<i>GRM6</i>	*	*	*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>GRP</i>	*	*	*	<i>ISM1</i>	*		*
<i>GSC2</i>	*	*	*	<i>ITGA8</i>	*		*
<i>GUCY2D</i>	*		*	<i>ITGA9</i>			*
<i>HAR1B</i>			*	<i>ITPKA</i>			*
<i>HBM</i>	*	*	*	<i>KBTBD11</i>			*
<i>HCN4</i>	*	*	*	<i>KCNA5</i>	*	*	*
<i>HDDC3</i>			*	<i>KCNA6</i>	*	*	*
<i>HES3</i>	*	*	*	<i>KCNC1</i>	*	*	*
<i>HES4</i>			*	<i>KCNK10</i>	*		*
<i>HES5</i>	*	*	*	<i>KCNK4</i>	*	*	*
<i>HES6</i>			*	<i>KCNN1</i>		*	
<i>HMGA2</i>	*	*	*	<i>KCNN2</i>	*	*	*
<i>HMX1</i>	*		*	<i>KCNQ2</i>	*	*	*
<i>HNF1B</i>	*	*	*	<i>KCNQ3</i>			*
<i>HOMER3</i>	*			<i>KCNQ4</i>	*		*
<i>HOOK1</i>	*	*	*	<i>KCNS2</i>	*	*	*
<i>HOXB4</i>	*	*	*	<i>KCNV1</i>	*	*	*
<i>HOXB7</i>	*	*	*	<i>KCTD8</i>	*	*	*
<i>HOXC12</i>	*	*	*	<i>KDF1</i>	*	*	*
<i>HOXC13</i>	*	*	*	<i>KIF21A</i>	*	*	*
<i>HOXD-AS2</i>	*	*	*	<i>KIF5C</i>		*	*
<i>HOXD4</i>	*	*	*	<i>KIRREL3</i>	*	*	*
<i>HOXD8</i>	*			<i>KLF4</i>	*	*	*
<i>HS3ST5</i>	*	*	*	<i>KLHL30</i>	*		*
<i>HTR1E</i>	*	*	*	<i>KLRG2</i>	*		*
<i>ID4</i>	*	*	*	<i>KRT18P55</i>	*		*
<i>IFITM10</i>			*	<i>KRTCAP3</i>	*	*	*
<i>IGF2R</i>			*	<i>LAG3</i>	*	*	*
<i>IGFBP3</i>			*	<i>LAMA1</i>	*	*	*
<i>IGFBP5</i>		*		<i>LAPTM4B</i>		*	
<i>IGFBPL1</i>	*	*	*	<i>LARGE1</i>	*	*	*
<i>IGSF9B</i>			*	<i>LARGE2</i>	*		*
<i>INCA1</i>			*	<i>LGR6</i>	*	*	*
<i>INHBB</i>	*	*	*	<i>LHB</i>			*
<i>INPP5A</i>			*	<i>LHFPL4</i>	*	*	*
<i>INSM2</i>	*	*	*	<i>LHX3</i>	*	*	*
<i>INSYN2A</i>	*	*	*	<i>LIN28B-AS1</i>			*
<i>IRF4</i>	*	*	*	<i>LINC00273</i>	*	*	*
<i>IRF5</i>			*	<i>LINC00667</i>			*
<i>IRX2</i>	*		*	<i>LINC00839</i>	*	*	*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>LINC00954</i>			*	<i>MIR3131</i>	*		*
<i>LINC01124</i>	*	*	*	<i>MIR4692</i>	*	*	*
<i>LINC02495</i>	*	*	*	<i>MIR9-1</i>	*	*	*
<i>LIPG</i>	*	*	*	<i>MKX</i>	*	*	*
<i>LMF1</i>	*	*	*	<i>MLPH</i>	*	*	*
<i>LMO1</i>	*	*	*	<i>MMD2</i>	*	*	*
<i>LOC155060</i>		*		<i>MMP21</i>			*
<i>LOC729970</i>			*	<i>MN1</i>	*	*	*
<i>LONRF2</i>			*	<i>MNX1-AS1</i>	*	*	*
<i>LPAR2</i>			*	<i>MOCS1</i>	*	*	*
<i>LRAT</i>	*	*	*	<i>MOXD1</i>	*	*	*
<i>LRP6</i>		*		<i>MPPED2</i>			*
<i>LRRC1</i>			*	<i>MSC</i>	*	*	*
<i>LRRC14B</i>	*		*	<i>MTA3</i>	*		*
<i>LRRC24</i>	*		*	<i>MTMR7</i>	*		*
<i>LRRC73</i>	*		*	<i>MYEF2</i>	*	*	*
<i>LTBP2</i>	*	*	*	<i>MYH14</i>	*	*	*
<i>LTBP3</i>	*	*	*	<i>MYO3A</i>	*	*	*
<i>LY6H</i>	*		*	<i>NACC2</i>	*	*	*
<i>M1AP</i>	*		*	<i>NALCN</i>	*	*	*
<i>MADCAM1</i>	*	*	*	<i>NAXD</i>	*		*
<i>MAFA</i>	*	*	*	<i>NBL1</i>	*		*
<i>MAFB</i>	*	*	*	<i>NCKAP1</i>	*	*	*
<i>MAMSTR</i>	*		*	<i>NECAB2</i>	*	*	*
<i>MAP6</i>	*	*	*	<i>NEUROG1</i>	*	*	*
<i>MAPK4</i>	*		*	<i>NFASC</i>	*	*	*
<i>MARCH11</i>	*	*	*	<i>NGF</i>	*	*	*
<i>MARCH9</i>	*	*		<i>NINL</i>	*	*	*
<i>MARK1</i>			*	<i>NKAIN4</i>	*	*	*
<i>MARVELD2</i>	*	*	*	<i>NKX1-2</i>	*	*	*
<i>MCOLN3</i>	*	*		<i>NKX2-3</i>	*		
<i>MEDAG</i>	*	*	*	<i>NKX2-8</i>	*	*	*
<i>MEG3</i>	*		*	<i>NKX3-2</i>	*	*	*
<i>MEIS1-AS3</i>	*	*		<i>NKX6-1</i>	*	*	*
<i>METRNL</i>			*	<i>NKX6-2</i>	*	*	*
<i>MFHAS1</i>	*		*	<i>NMRK2</i>	*	*	*
<i>MFSD6L</i>	*		*	<i>NOTO</i>	*	*	*
<i>MGC2889</i>	*	*	*	<i>NOTUM</i>	*	*	*
<i>MIR181D</i>	*		*	<i>NPAS1</i>	*		*
<i>MIR203A</i>	*		*	<i>NPAS2</i>	*	*	*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>NPTX1</i>	*	*	*	<i>PLEKHG1</i>	*	*	*
<i>NPW</i>	*	*	*	<i>PLEKHG6</i>	*	*	*
<i>NRARP</i>			*	<i>PLLP</i>	*		*
<i>NRN1</i>	*	*	*	<i>PMEPA1</i>	*		*
<i>NSD1</i>	*	*	*	<i>PNPLA5</i>	*	*	*
<i>NUP210</i>	*	*	*	<i>PODXL</i>	*	*	
<i>NXPH4</i>	*	*	*	<i>PODXL2</i>			*
<i>OLFM2</i>	*			<i>POGLUT2</i>			*
<i>ONECUT3</i>	*	*	*	<i>PP7080</i>	*		*
<i>OPRK1</i>	*	*	*	<i>PPL</i>	*	*	*
<i>OSBP2</i>	*	*	*	<i>PPP1R1B</i>	*	*	*
<i>OSCP1</i>	*	*	*	<i>PPP2CB</i>			*
<i>OVOL1</i>	*		*	<i>PPP2R2C</i>	*		
<i>OXT</i>	*	*	*	<i>PRDM16</i>	*	*	*
<i>PALM</i>			*	<i>PRDX2</i>	*		*
<i>PAPLN</i>	*	*	*	<i>PRKD2</i>			*
<i>PAPPA</i>	*	*	*	<i>PRKG2</i>	*		
<i>PARD3B</i>	*	*	*	<i>PROB1</i>	*	*	*
<i>PAWR</i>		*		<i>PROKR2</i>	*	*	*
<i>PAX3</i>	*	*	*	<i>PROSER3</i>	*	*	*
<i>PCSK2</i>	*	*	*	<i>PROX1</i>	*	*	*
<i>PDE10A</i>	*		*	<i>PRR35</i>	*	*	*
<i>PDE4DIP</i>	*	*	*	<i>PRRX2</i>	*	*	*
<i>PDGFA</i>	*	*	*	<i>PRTG</i>	*		
<i>PDLIM4</i>			*	<i>PSD</i>			*
<i>PDX1</i>	*	*	*	<i>PTF1A</i>	*	*	*
<i>PDZD2</i>	*	*	*	<i>PTGDR</i>	*	*	*
<i>PEAR1</i>	*	*	*	<i>PTK2</i>	*	*	*
<i>PENK</i>	*	*	*	<i>PTPN3</i>	*	*	*
<i>PFN3</i>	*		*	<i>PTPRS</i>	*	*	*
<i>PGLYRP1</i>	*	*	*	<i>PUSL1</i>			*
<i>PGR</i>	*		*	<i>RAB6B</i>			*
<i>PIAS4</i>	*	*		<i>RAC3</i>			*
<i>PIF1</i>			*	<i>RAD21L1</i>		*	*
<i>PITPNM2</i>			*	<i>RADIL</i>	*		*
<i>PKMYT1</i>	*	*		<i>RAI1</i>			*
<i>PKNOX2</i>			*	<i>RAMP3</i>	*	*	*
<i>PLAC9</i>			*	<i>RBM24</i>	*	*	*
<i>PLCL1</i>	*			<i>REEP6</i>	*		*
<i>PLEKHD1</i>	*	*	*	<i>RELN</i>	*		*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>RET</i>	*			<i>SLC26A5</i>		*	*
<i>RFX6</i>	*	*	*	<i>SLC35F1</i>	*	*	*
<i>RGS10</i>			*	<i>SLC35F3</i>	*	*	*
<i>RGS7</i>	*	*	*	<i>SLC35G1</i>			*
<i>RIMBP2</i>	*	*	*	<i>SLC37A3</i>			*
<i>RMDN2</i>	*	*	*	<i>SLC41A1</i>	*	*	*
<i>RNF207</i>	*		*	<i>SLC49A3</i>	*		*
<i>RNF225</i>	*	*	*	<i>SLC4A11</i>	*		*
<i>RORB</i>	*	*	*	<i>SLC6A11</i>	*	*	*
<i>RPH3A</i>	*	*	*	<i>SLC6A5</i>	*	*	*
<i>RPRM</i>	*	*	*	<i>SLC7A14</i>	*	*	*
<i>RRAD</i>	*		*	<i>SMIM1</i>	*	*	*
<i>RSPO1</i>			*	<i>SMOC2</i>	*	*	*
<i>RSPO4</i>	*	*	*	<i>SMTNL2</i>	*	*	*
<i>S1PR5</i>	*	*	*	<i>SOC33</i>	*		*
<i>SALL3</i>	*	*	*	<i>SORBS3</i>	*	*	*
<i>SAMD5</i>			*	<i>SORT1</i>			*
<i>SBSPON</i>	*	*	*	<i>SOWAHC</i>	*	*	*
<i>SCARF2</i>	*	*	*	<i>SOX11</i>	*		*
<i>SCRIB</i>	*			<i>SOX14</i>	*	*	*
<i>SCRN1</i>	*	*	*	<i>SP8</i>	*	*	*
<i>SCRT1</i>			*	<i>SPEG</i>	*	*	
<i>SCTR</i>	*	*	*	<i>SQSTM1</i>	*		*
<i>SCUBE1</i>	*	*	*	<i>SREBF1</i>	*		*
<i>SDC3</i>	*	*	*	<i>ST6GALNAC2</i>			*
<i>SDR16C5</i>	*	*	*	<i>ST8SIA2</i>	*	*	*
<i>SEL1L3</i>	*	*	*	<i>ST8SIA5</i>	*	*	*
<i>SFMBT1</i>			*	<i>STAC2</i>	*	*	*
<i>SGCZ</i>	*	*	*	<i>STK25</i>			*
<i>SH3GL2</i>	*	*	*	<i>STK32A</i>			*
<i>SHC3</i>	*	*	*	<i>STMND1</i>	*	*	*
<i>SHISA6</i>	*	*	*	<i>SULT4A1</i>	*	*	*
<i>SHLD2</i>			*	<i>SYNE2</i>			*
<i>SLC13A5</i>	*		*	<i>SYT12</i>	*	*	*
<i>SLC15A1</i>	*	*	*	<i>SYT17</i>			*
<i>SLC18A2</i>	*			<i>SYT5</i>	*		*
<i>SLC19A1</i>	*	*		<i>TAF5A</i>	*	*	*
<i>SLC22A23</i>	*	*	*	<i>TBX15</i>	*	*	*
<i>SLC24A4</i>	*	*	*	<i>TBX4</i>	*	*	*
<i>SLC25A21</i>	*	*	*	<i>TCEA3</i>	*	*	*

Table 2 (Continued).

Gene	dMML	dNME	dPDM	Gene	dMML	dNME	dPDM
<i>TCF7L1</i>	*			<i>VAX1</i>	*	*	*
<i>TCL1A</i>	*	*	*	<i>VENTX</i>	*	*	*
<i>TDRP</i>	*	*	*	<i>VPS28</i>	*		
<i>TEAD3</i>			*	<i>VPS37D</i>	*	*	*
<i>TEDC2</i>			*	<i>VSIG10L</i>			*
<i>TENM3-AS1</i>			*	<i>VSX1</i>	*	*	*
<i>THSD7B</i>	*	*	*	<i>VSX2</i>	*	*	*
<i>THY1</i>	*	*	*	<i>VTN</i>	*	*	*
<i>TIAM1</i>	*	*	*	<i>VWA5B2</i>	*	*	*
<i>TLX3</i>	*	*	*	<i>VWC2</i>	*	*	*
<i>TMEM130</i>			*	<i>WASF3</i>	*	*	*
<i>TMEM132D</i>	*	*	*	<i>WFS1</i>	*	*	*
<i>TMEM151A</i>	*		*	<i>WI2-2373I1.2</i>	*	*	*
<i>TMEM178B</i>	*	*	*	<i>WIPF3</i>	*	*	*
<i>TMEM200C</i>	*	*	*	<i>WNK2</i>	*	*	*
<i>TMEM30B</i>	*	*	*	<i>WNT6</i>			*
<i>TMIE</i>	*	*	*	<i>WNT9B</i>	*	*	*
<i>TPSD1</i>	*	*	*	<i>WSCD2</i>	*	*	*
<i>TRIM67</i>	*	*	*	<i>XKR4</i>	*	*	*
<i>TRIO</i>	*			<i>XKR6</i>	*	*	*
<i>TRPC4</i>	*	*	*	<i>YBX3P1</i>	*		
<i>TSSK3</i>	*	*	*	<i>ZAR1</i>	*	*	*
<i>TST</i>		*		<i>ZBTB42</i>			*
<i>TTC34</i>	*	*	*	<i>ZFP2</i>	*	*	*
<i>TUB</i>	*	*	*	<i>ZFP42</i>	*		*
<i>TULP1</i>	*		*	<i>ZMYND15</i>			*
<i>TUNAR</i>	*	*	*	<i>ZNF503</i>		*	
<i>TWIST2</i>	*	*	*	<i>ZNF606</i>	*	*	*
<i>UCHL1</i>			*	<i>ZNF703</i>	*	*	*
<i>UNC5A</i>	*	*	*	<i>ZNF710</i>	*	*	*
<i>UTF1</i>	*	*	*	<i>ZNF746</i>			*
<i>VAV2</i>			*	<i>ZNF831</i>	*		

Table 3. Gene set enrichment analysis (GSEA) results (overlaps with the 'GO gene sets' in MSigDB) of the 574 genes in Table 2 exhibiting statistically significant PDM differences (dPDM genes). GSEA Collection: C5; overlaps shown: 25; # genes in collection: 10,192; # genes in comparison: 570; # genes in universe: 38,404.

Gene set name	Genes in gene set	Genes in overlap	FDR Q-value
GO_NEUROGENESIS	1625	108	7.62E-36
GO_NEURON_DIFFERENTIATION	1365	93	2.57E-31
GO_CELL_CELL_SIGNALING	1665	98	3.07E-28
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY_RNA_POLYMERASE_II_SPECIFIC	1094	79	5.28E-28
GO_SEQUENCE_SPECIFIC_DNA_BINDING	1189	79	1.11E-25
GO_NEURON_PROJECTION	1317	83	1.22E-25
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	1334	81	6.14E-24
GO_CHROMATIN	1226	77	1.37E-23
GO_TRANSCRIPTION_REGULATOR_ACTIVITY	1718	90	1.85E-22
GO_NUCLEAR_CHROMOSOME	1273	76	5.61E-22
GO_ANIMAL_ORGAN_MORPHOGENESIS	1071	69	1.27E-21
GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	928	62	4.89E-20
GO_NEURON_DEVELOPMENT	1114	68	4.89E-20
GO_SYNAPSE	1332	74	1.27E-19
GO_REGULATORY_REGION_NUCLEIC_ACID_BINDING	960	62	2.24E-19
GO_REGULATION_OF_CELL_DIFFERENTIATION	1881	88	7.21E-19
GO_SOMATODENDRITIC_COMPARTMENT	834	57	7.21E-19
GO_CELLULAR_COMPONENT_MORPHOGENESIS	1141	66	2.81E-18
GO_CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT	995	61	5.29E-18
GO_POSITIVE_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_II	1198	67	7.46E-18
GO_PATTERN_SPECIFICATION_PROCESS	446	41	9.31E-18
GO_SEQUENCE_SPECIFIC_DOUBLE_STRANDED_DNA_BINDING	889	57	1.15E-17
GO_CELL_PROJECTION_ORGANIZATION	1545	76	2.45E-17
GO_CHROMOSOME	1730	81	2.45E-17
GO_EMBRYONIC_MORPHOGENESIS	592	46	3.16E-17