MethylBERT: A Transformer-based model for read-level DNA methylation pattern identification and tumour deconvolution

Yunhee Jeong[1], Karl Rohr[2], Pavlo Lutsik[1,3]

[1]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, Germany

[2] Biomedical Computer Vision Group, BioQuant, IPMB, Heidelberg University, Im Neuenheimer Feld 267, Heidelberg, Germany

[3] Department of Oncology, Catholic University (KU) Leuven, Leuven, Belgium

Abstract

DNA methylation (DNAm) is a key epigenetic mark that shows profound alterations in cancer. Although read-level methylomes enable more in-depth DNAm analysis due to the broad coverage and preservation of rare cell-type signals, the majority of published DNAm analysis methods have targeted array-based data such as EPIC/450K array. Here, we propose MethylBERT, a novel Transformer-based read-level methylation pattern classification model. MethylBERT identifies tumour-derived sequence reads based on their methylation patterns and genomic sequence using a Bidirectional Encoder Representations from Transformers (BERT) model. Based on the calculated classification probability, the method estimates tumour cell fractions within bulk samples and provides an assessment of the model precision. In our evaluation, MethylBERT outperforms existing deconvolution methods and demonstrates high accuracy regardless of methylation pattern complexity, read length and read coverage. Moreover, we show its potential for accurate non-invasive early cancer diagnostics by applying MethylBERT to liquid biopsy samples collected from cancer patients. MethylBERT represents a significant advancement in read-level methylome analysis. It will increase the accuracy of tumour deconvolution and enhance circulating tumour DNA studies.

Introduction

DNA methylation (DNAm) refers to a methyl group added to nucleotide bases in the DNA. Aberrant methylation patterns at CpGs reflect epigenetic heterogeneity in tumours [1]. Therefore, DNAm data has been broadly used for estimating tumour purity to examine the tumour epigenetic landscape relevant to clinical outcomes, tumour diagnosis and phenotypic

characteristics [2–4]. Compared to DNA mutations which are often scarce in many cancer types at early stages and could also be detected in healthy samples [5], DNAm has great benefits as a tumour biomarker. Assorted experimental methods can be applied to disentangle cell-type composition of bulk tumours such as single-cell and spatial omics technologies or flow cytometry. However, computational tumour purity estimation, also known as tumour deconvolution, is more cost-efficient and avoids many technical biases inherent to experimental approaches. It is also useful for numerous biobanked DNAm data since their biological samples are not available any more.

DNAm can be profiled by sequencing methods such as whole genome bisulfite sequencing (WGBS) or reduced representation bisulfite sequencing (RRBS), as well as microarray methods like Infinium 450K/EPIC array. High-quality sequencing-based profiling delivers sequence reads over a broad range of genomic coverage and preserves single-molecule signals of rare cell populations [6]. Using sequencing-based data for tumour deconvolution is even more crucial in circulating tumour DNA (ctDNA) analysis. ctDNA analysis facilitates non-invasive early diagnosis, prognosis and treatment response monitoring for cancer patients [7–9]. However, the very low quantity of ctDNA commonly found in liquid biopsy (< 5%) makes the analyses challenging. In this case, sequencing-based DNAm tumour deconvolution is an ideal option to detect rare tumour signals from the ctDNA. Nevertheless, most tumour deconvolution methods have been developed for array-based DNAm profiles because of the intuitive application of classical linear algebraic algorithms to matrices of DNA methylation beta-values. Moreover, our analysis showed that existing sequencing-based deconvolution methods do not perform better than array-based methods implying that they fail to fully exploit the advantages of sequencing data for accurate inference [10].

To overcome these limitations, we propose MethylBERT, a novel deep learning method for read-level methylation pattern identification and tumour deconvolution based on Bidirectional Encoder Representations of Transformers (BERT). MethylBERT uses a pre-trained BERT model to encode read-level methylomes and classifies sequencing reads into tumour or normal cell types. The estimated posterior probability of cell types is used for tumour purity estimation by applying Bayes' theorem and maximum likelihood estimation. Along with purity estimation, MethylBERT provides the model precision based on the Fisher information and estimation adjustment taking the region-wise tumour purity into account. To our knowledge, MethylBERT is the first Transformer-based deep learning method for tumour deconvolution using sequencing-based DNAm data (Supplementary Table 1). PRISM and MethylPurify employ a classical Expectation-Maximisation (EM) algorithm to estimate tumour purity from sequencing-based DNAm profiles [11,12], whereas the only deep learning method, DISMIR, is based on a Long Short-Term Memory (LSTM) model [13], which heavily relies on the direction within sequences.

We have thoroughly evaluated MethylBERT and compared it with existing methods, which demonstrated the remarkable performance of MethylBERT in read-level methylation pattern classification and tumour deconvolution. In our evaluation, we not only investigate the performance of MethylBERT but also analyse what the model actually learns via pre-training using reference genome sequences. This has not been well-addressed in previous work on BERT-based methods for genomic sequence data. Moreover, we suggest using the Fisher information to estimate the precision of the deconvolution model. This gives guidance about

how accurate the estimated tumour purity is, which represents essential information for analysing bulk samples without ground-truth tumour purity. Finally, we also propose using MethylBERT as a non-invasive method for early diagnosis of cancers and other diseases based on ctDNA analysis.

Results

MethylBERT overview

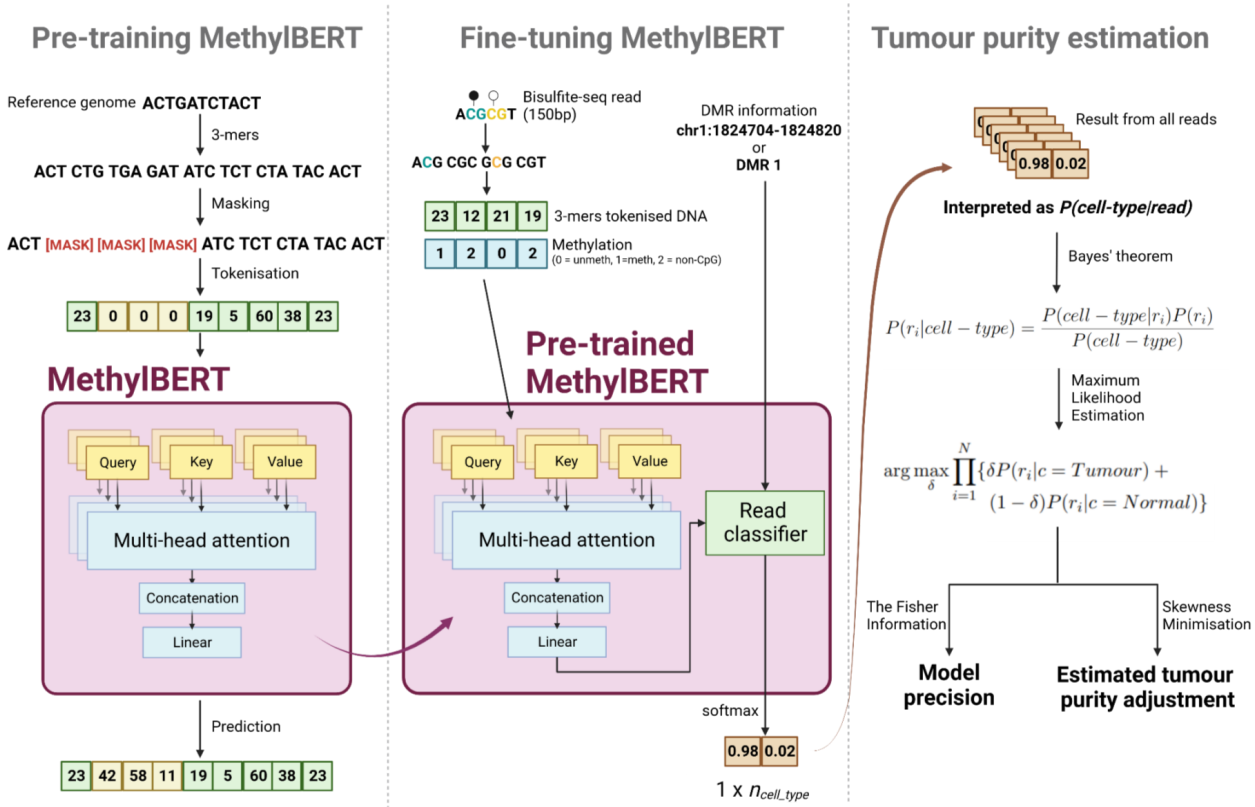

Figure 1. MethylBERT overview. The main three steps are separated by dotted lines.

The aim of MethylBERT is to classify sequence reads with their CpG methylation patterns and sequence into dichotomous cell-type-related classes, e.g., tumour or normal. Applied to a set of genomic regions, it also provides a global maximum likelihood estimate of cell-type proportions for the two classes, e.g., tumour purity in the tumour-normal deconvolution scenario.

The method includes three main steps (Figure 1). First, the MethylBERT model is pre-trained with a reference genome processed into 3-mers sequences. We use the Masked Language Model (MLM) for pre-training while the original BERT method performs pre-training of the model via both MLM and next sentence prediction (NSP) [14]. NSP is known to make the BERT model understand semantic dependencies across sentences [15], however, since we focus on the methylation pattern of unpaired reads, NSP was disregarded for MethylBERT

pre-training. The overall pre-training scheme to use 3-mers for MLM is inspired by DNABERT [16].

After pre-training, the MethylBERT model is fine-tuned to learn the read-level methylation pattern classification task (Figure 1 middle). The encoder network in MethylBERT encodes a read-level reference DNA sequence and CpG methylation patterns (Supplementary Figure 1). Although reference DNA sequences do not contain any tumour-specific genetic information, they carry the exact position of a read in a given region as well as sequence features associated with specific methylation patterns. This information is necessary for the model to learn which CpGs are more likely to show tumour-specific signals. Previous studies also have pointed out that integrated information of DNA sequences and methylation can improve modelling [13,17]. The encoded vector from the encoder network is concatenated with an embedded DMR label. The DMR label is provided as information to determine region-wise tumour-specific methylome profile. Finally, a cell-type classifier network following the encoder network computes a 2-dimensional vector. The final output converted with a softmax function is interpreted as the posterior probability $P(c_j|r_i)$ of the cell type c_j given a read r_i. In our case, the cell type is either tumour (*T*) or normal (*N*). Classification is performed by assigning the cell type with a higher posterior probability.

For the final tumour purity estimation, we apply Bayes' theorem to compute the posterior probability $P(r_i|c_j)$ using $P(c_j|r_i)$ assuming that every read has an equal marginal probability. The prior probability of the cell types $P(c_j)$ is calculated by taking the ratio of the number of reads from cell type c_j over the total reads in the training dataset. Afterwards, the tumour purity is determined by maximum likelihood estimation. The estimated tumour purity can also be adjusted based on the skewness of the region-wise tumour ratio. The adjustment is particularly useful when the analysed bulks have a very high or low ratio of tumour-derived reads. A detailed description of the three steps of MethylBERT is provided in Methods.

MethylBERT classifies complex read-level methylation patterns

Complex methylation patterns in tumours influence oncogene and tumour suppressor gene regulations and contribute to tumour heterogeneity [18,19]. Therefore, a read-level tumour deconvolution model should be capable of handling the complex methylation patterns. We simulated read-level methylomes in different scenarios to evaluate the robustness of the MethylBERT read classification with respect to the pattern complexity. Details about the simulation algorithm are described in Methods. We compared our method with CancerDetector [20] and DISMIR [13] as well as a baseline method implemented using a hidden Markov model (HMM) designed to classify read-level methylation patterns into cell types.

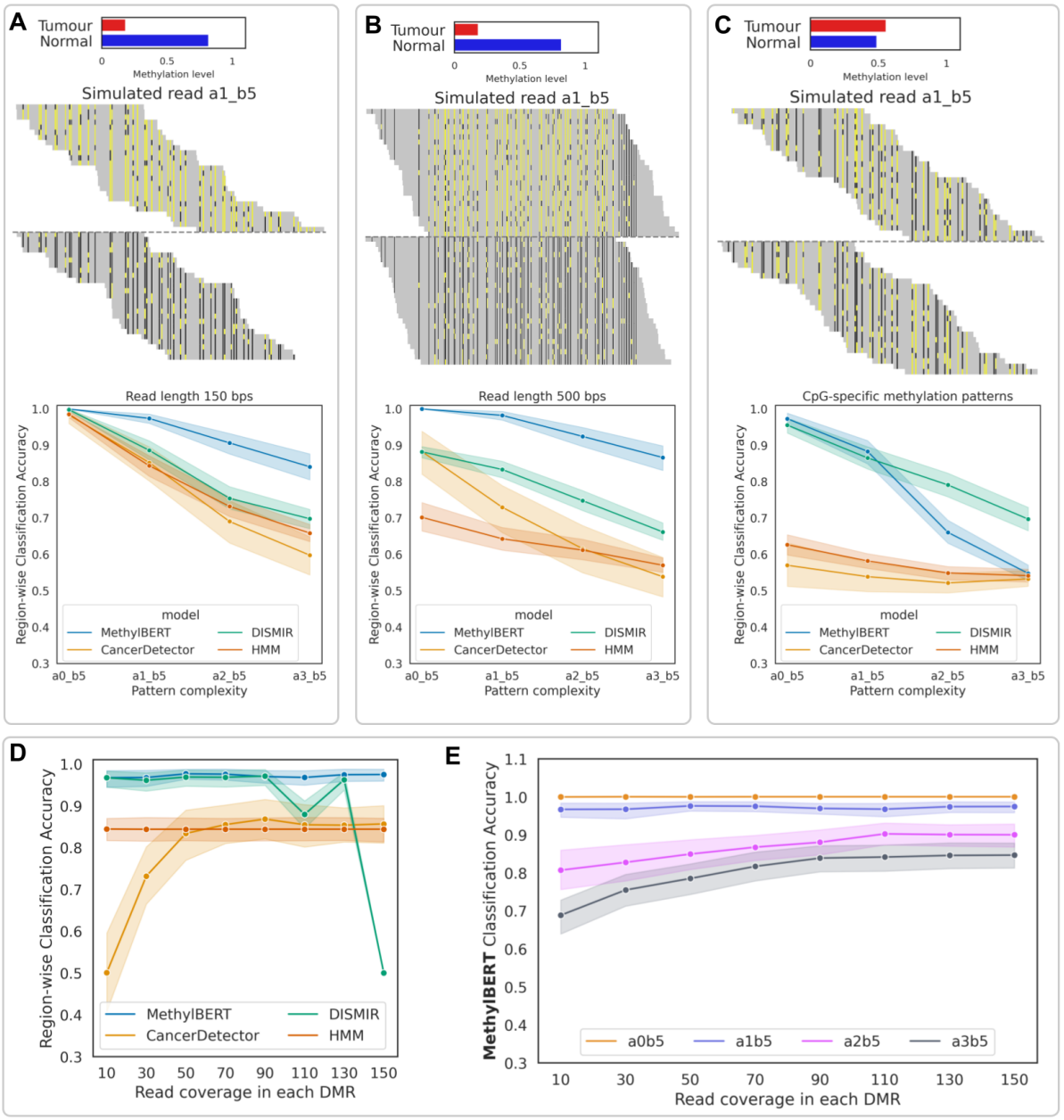

Figure 2. Read-level methylome classification results. ***A-C.*** *An example of sampled read-level methylation patterns for tumour and normal cell types and the classification accuracy comparison across different complexities (A) with read length 150 bp, (B) with read length 500 bp, and (c) with CpG-specific methylation patterns.* ***D.*** *Read-level methylome classification accuracy comparison for different read coverages.* ***E.*** *MethylBERT read-level methylome classification accuracy for different read coverages and pattern complexities.*

In our first scenario, we simulated different complexities of 150 bps read-level methylomes from a beta-binomial distribution. Our simulation algorithm generates more complex methylation patterns with an increasing α value of the beta distribution (Supplementary Figure 2A and B). For all complexities of methylation patterns, MethylBERT outperformed the other three methods in the classification task (Figure 2A). All four methods yielded the highest accuracy and the lowest deviation of accuracy values over DMRs for the simplest methylation patterns (a0_b5). Although the accuracy decreases with increasing complexity

for all methods, MethylBERT still classified reads more accurately than the other methods in each case. We further simulated longer read-level methylation patterns with 500 bp length. In this case, the reads cover both a DMR and non-DMRs on the genome resulting in more complicated methylation patterns, because most DMRs are shorter than 500 bps (Supplementary Figure 2C). With the 500bps of simulated reads, MethylBERT again performed better than the other methods regardless of the complexity (Figure 2B). It likewise achieved more precise classification for simpler complexity keeping the lowest deviation of accuracy at the complexity a0_b5. On the other hand, for the same complexity level, CancerDetector and HMM performed worse and had a higher deviation of accuracy in the results with 500 bp reads than in the results with 150 bp reads. Overall, the deep learning methods MethylBERT and DISMIR yielded better classification results for the 500 bp read simulation.

Neighbouring CpGs usually have a consistent methylation pattern, but the innate inclination of methylation for each CpG is a non-trivial factor in determining methylation patterns [21]. In addition, erroneous patterns differing from neighbouring CpGs are more commonly found in cancer [22]. Thus, we conducted the same evaluation for the extreme scenario where only CpG-specific methylation patterns differed between tumour and normal whereas region-specific methylation levels were almost identical (Methods, Supplementary Figure 2D). CancerDetector and HMM could not classify reads into cell types despite the clear pattern difference in the case of a0_b5 and a1_b5, whereas MethylBERT performed best with the exception of highly noisy patterns in a2_b5 and a3_b5 (Figure 2C). These results demonstrate that MethylBERT is capable of tumour-specific signals in CpG level not misled by the average methylation level in a region.

Read coverage is another important factor which can affect read-level methylome analysis. Many deconvolution algorithms pre-select regions based on the minimum number of CpGs and read coverage within the region so that sufficient methylomes can be used for deconvolution [10]. Hence, we simulated read-level methylomes with variable read coverage in DMRs. Again, MethylBERT achieved the best methylation pattern classification performance (Figure 2D). Although higher read coverage is usually considered better for DNAm analysis, we found that the performance of DISMIR drastically decreases with higher read coverage, possibly due to the higher variance of the methylation patterns (Supplementary Figure 2E). On the other hand, MethylBERT can handle highly varying methylation patterns. MethylBERT keeps a high accuracy regardless of the coverage for the complexity a0_b5 and a1_b5. However, the accuracy converges at the highest value when the read coverage is higher than 100 for the complexity a2_b5 and a3_b5 (Figure 2E).

MethylBERT pre-training allows the model to understand sequence features

Bidirectional pre-training is a pivotal feature of the BERT model [14], alleviating the restricted choice of a model architecture. It was shown by Ji et al. [16] that the pre-trained BERT model can be successfully fine-tuned for various DNA sequence analyses such as promoter region prediction. Clark et al. [23] have carefully described what kind of information BERT

learns during pre-training in NLP. Yet, the efficacy of BERT pre-training on DNA sequences is still poorly understood. Therefore, we have wondered how pre-training can affect the methylation pattern classification and what kind of context in DNA sequences BERT learns via pre-training.

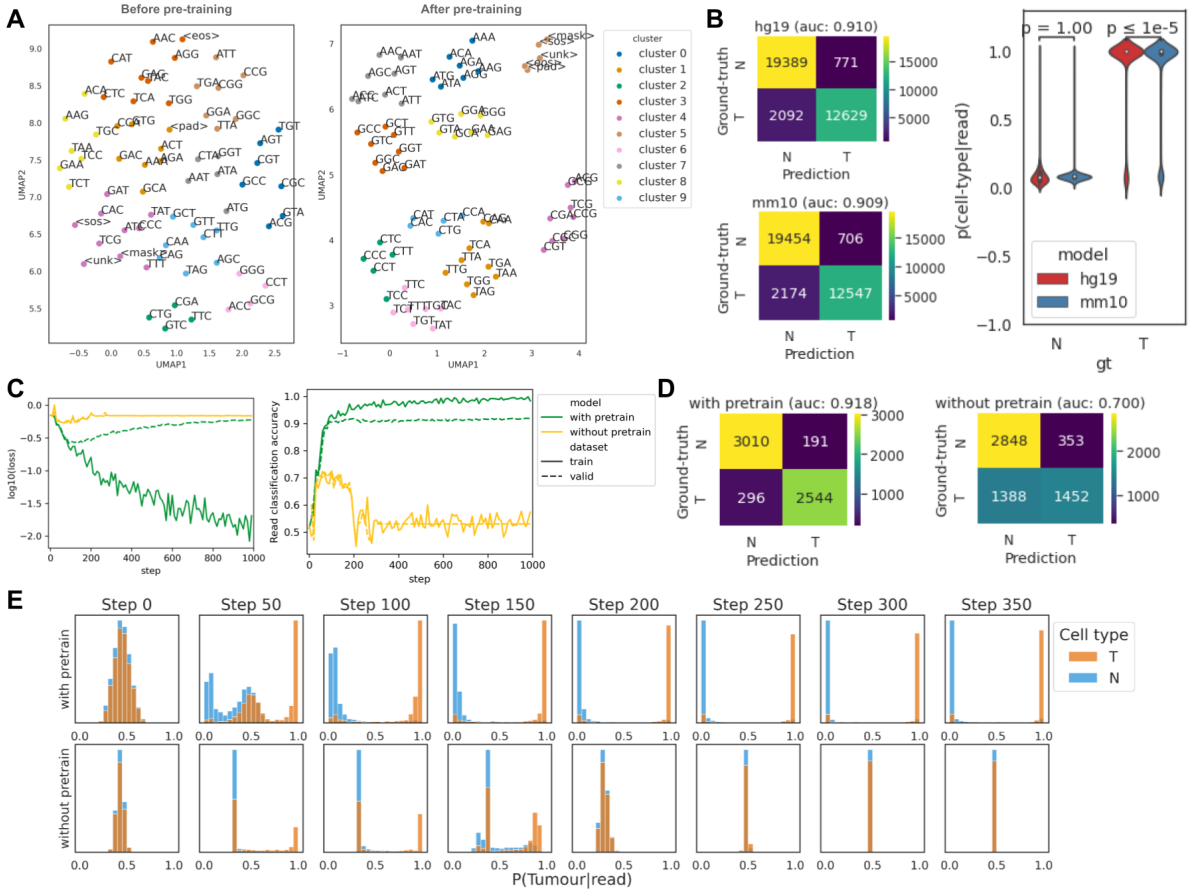

Figure 3. Impact of pre-training in MethylBERT. ***A.*** *UMAP plot of 3-mers token embeddings before and after pre-training. The clusters were made via k-means clustering.* ***B.*** *Confusion matrix of read classification results by the model pre-trained with human genome hg19 (top left) and with mouse genome mm10 (bottom left). Distribution of P(cell type=Tumour|read) in both cell types calculated by the two pre-trained models (right). P-values in the violin plot were calculated using paired t-test statistics.* ***C.*** *Training (solid line) and validation (dotted line) curves of MethylBERT with and without pre-training (green and yellow). Both graphs are plotted every 10 steps.* ***D.*** *Confusion matrix of read classification results by the MethylBERT model with and without pre-training calculated at the step when each model achieved the best validation performance.* ***E.*** *Histogram of P(cell type=Tumour|read) in tumour and normal reads (orange and blue each) calculated by MethylBERT with and without pre-training (top and bottom).*

We have primarily found that pre-training enables the BERT model to understand the mutual relationships between DNA 3-mers (Figure 3A). Even though any information about CpGs or methylation has not been provided, BERT distinguishes 3-mers tokens including “CG” from other tokens (Cluster 4). This might be driven by the repetitive “CG” patterns especially occurring in CpG islands. In addition, the pre-trained BERT model is able to associate paired DNA nucleotides with each other (A-T and C-G). UMAP2 embedding divides the 3mers tokens into two groups: those that start with C/T and the others starting with A/G. Each cluster is composed of tokens whose first nucleotide is the same and whose last nucleotide

makes a nucleotide pair. For example, in cluster 7, all tokens start with A and end with C/T. We hypothesise that BERT pre-training identified the nucleotide pairs because of Chargaff's second parity rule that the amount of two nucleotides in a pair is approximately equal in a single DNA strand. Special tokens (e.g. <unk> or <mask>) also make up a separate cluster from the other DNA 3-mers tokens only after pre-training.

We have also evaluated the influence of pre-training on the identification of tumour methylation patterns. For this, we compared the performance of read classification when the MethylBERT model was pre-trained and when it was not pre-trained. The comparison was done by utilising diffuse large B cell lymphoma (DLBCL) and non-neoplastic B cell samples. In order to avoid the additional influence carried by the dominant promoter hypermethylation in the tumour, we selected 50 DMRs where tumour cells are hypermethylated and hypomethylated, respectively (Supplementary Figure 3A and B). Both MethylBERT models with and without pre-training gradually decrease the loss value during the early steps of fine-tuning (Figure 3C). However, when MethylBERT is not pre-trained, the loss value increases again after 100 steps and the accuracy eventually converges around 0.5. The confusion matrix of classified reads also shows that pre-trained MethylBERT achieves far more accurate classification results (Figure 3D).

Figure 3E shows the change in the probability $P(cell\ type\ =\ T|read)$ during MethylBERT fine-tuning. When MethylBERT is pre-trained with the reference genome, the probability distribution of tumour reads and normal reads start separating from each other at step 50. Over further steps, the model yields higher probability values of correct cell types for the reads. For instance, at step 350, $P(cell\ type\ =\ T|read)$ of normal reads are close to 0 and the probability of tumour reads are close to 1. On the other hand, MethylBERT without pre-training could distinguish some tumour reads from normal reads until step 150, but afterwards the $P(cell\ type\ = Tumour|read)$ distribution of two cell types becomes indistinguishable. According to the correlation between the estimated probability and methylation level of reads, both models classify the reads mainly based on the methylation level at step 50 (Supplementary Figure 3C). However, only the pre-trained model can overcome the bias and, in the following steps, keep the accuracy high without $P(cell\ type\ = Tumour|read)$ not being strongly correlated with the methylation level.

Pre-training on an entire reference genome requires a long training time, so the cross-species applicability of the pre-trained MethylBERT model will augment the utility to analyse various samples. For this reason, we have investigated the discrepancy in fine-tuning performance between pre-trained models with human (hg19) and mouse (mm10) genomes. The classification results of DLBCL and non-neoplastic B cell reads show that both human and mouse genomes are eligible as pre-training data for human cancer analysis (Figure 3B). The read classification area under the curve (AUC) score calculated in the validation set shows a difference below 0.001 between the two reference genomes. The distribution of calculated $P(cell\ type\ =\ Tumour|read)$ also does not significantly differ in non-neoplastic B cell reads. However, in the DLBCL reads, the two sets of probability values still have a p-value below $1.0\ \times\ 10^{-5}$ for paired t-test statistics.

From the results in this section, we claim that pre-training is a vital step enabling the MethylBERT model to understand the major sequence features of the DNA. Furthermore,

the MethylBERT fine-tuning can prevent bias towards read-wise methylation levels only when pre-training is performed. Finally, the cross-species applicability of the pre-training broadens the range of available samples for MethylBERT analysis.

MethylBERT deconvolutes mixed tumour methylomes and accurately estimates tumour purity

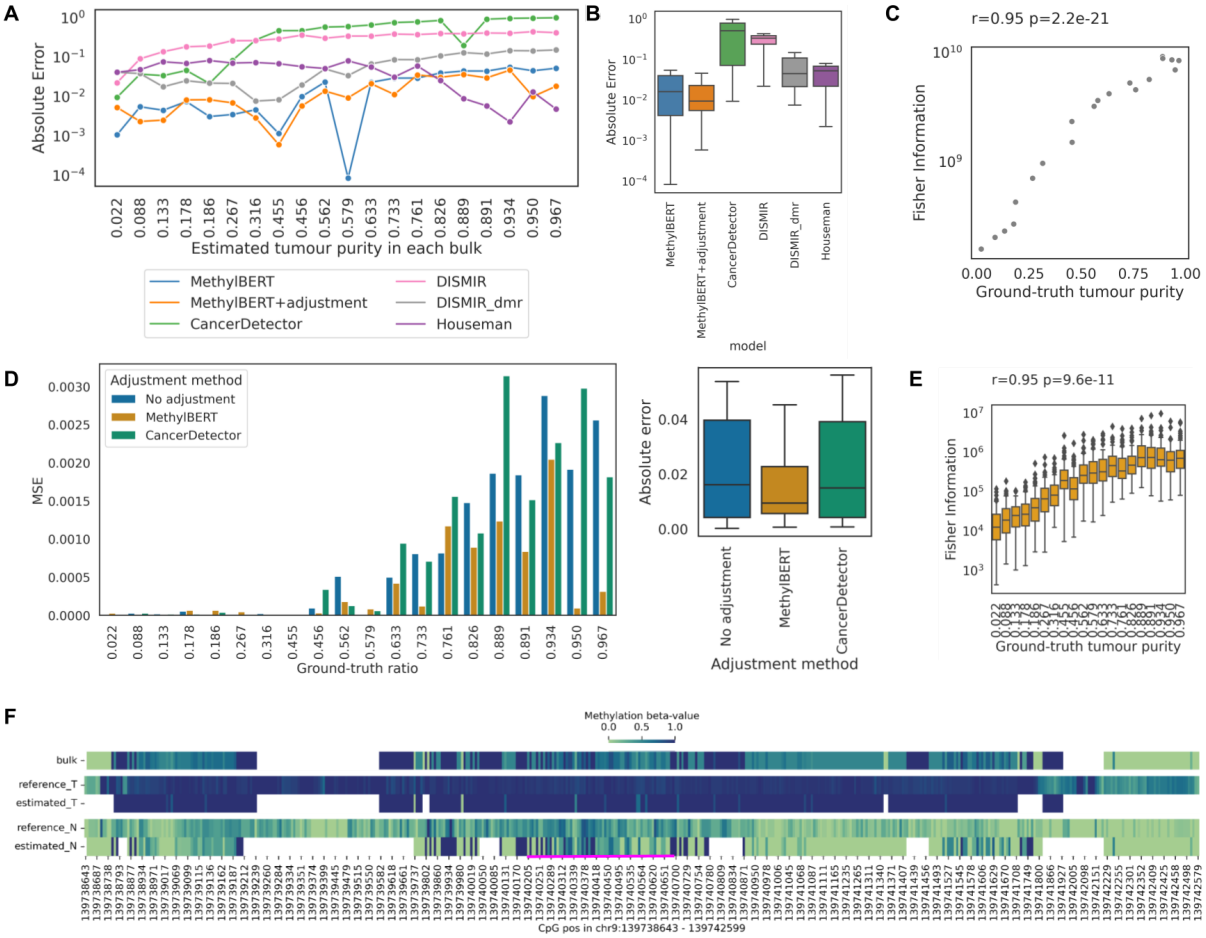

Figure 4. Tumour deconvolution and the estimation adjustment results for DLBCL pseudo-bulk samples. ***A-B.*** *Absolute error between the ground-truth and estimated tumour purity. Bulks are sorted according to the tumour purity on the x-axis in the line plot (A). In the box plot, the middle bar, bottom edge and top edge of the box plot indicate the median, first quantile and third quantile values, respectively (B).* ***C.*** *Correlation between the ground-truth tumour purity and calculated Fisher information values during maximum likelihood estimation.* ***D.*** *Performance comparison of different estimation adjustment methods. The histogram shows the mean squared error between the ground-truth and the estimated tumour purity for pseudo-bulks sorted by the ground-truth (left). The boxplot presents the absolute error (right).* ***E.*** *Calculated region-wise Fisher information for each pseudo-bulk when the MethylBERT estimation adjustment was applied.* ***F.*** *Estimated CpG-wise methylation beta values of bulk 8 within DMR 14 (chr9:139738643-139742599) compared with the reference tumour and normal methylation patterns (bottom four rows). The first row shows the original beta values of bulk 8 in DMR 14. The complex methylation patterns in the normal cell type (chr9:139740205-139740700) are indicated by the magenta line.*

We next evaluated the tumour purity estimation performance of MethylBERT using *in silico* generated pseudo-bulk samples. The samples were obtained by mixing reads randomly sampled from DLBCL and non-neoplastic B cell samples with controlled proportions. For the comparison, we benchmarked against CancerDetector [20], DISMIR [13] and Houseman's method [24]. Houseman's method performed best in diverse deconvolution experiments in our previous benchmarking study [10]. Although DISMIR features a procedure for informative genomic region selection for deconvolution, it could not make reasonable estimates with self-selected regions. Therefore, we applied DISMIR on the same DMRs used for the other methods and this is referred to as 'DISMIR_dmr' in the following. For the MethylBERT performance, we present the performance from both models with and without estimation adjustment (Methods) to study the impact of the adjustment.

MethylBERT outperformed other methods with respect to the absolute error between the ground-truth and estimated tumour ratio (Figure 4A and B). Although Houseman's method achieved better performance for the bulks with a high tumour ratio, it could not accurately estimate the purity when the ground-truth tumour ratio is lower. On the contrary, CancerDetector performed better for the bulk samples where tumour fraction was minor. However, MethylBERT maintained its high accuracy for both low and high tumour purities.

As a measure of model estimates certainty, MethylBERT provides the precision of the estimated tumour purity using Fisher information. When multiple samples are dissected by MethylBERT, the user can obtain information on which sample is more likely to have an accurate estimate by comparing the Fisher information indicating the precision of likelihood estimation [25]. The Fisher information calculated over log-likelihood values positively correlates with the ground-truth tumour purity in the pseudo-bulks (p-value = 2.2×10^{-21}) (Figure 4C). Given that the read classification accuracy tends to be lower with higher tumour purity in pseudo-bulk samples (Supplementary Figure 4), we confirm that the Fisher information can offer a comparable measurement of the tumour purity estimation precision.

To further improve the accuracy of tumour purity estimation, MethylBERT employs estimation adjustment based on the distribution of region-wise (local) estimated tumour purity (Methods). This is conceptually equal to the 'removal of confounding factors' in CancerDetector [20], as it also copes with outlier regions where the estimated ratio of tumour-derived reads is different from the majority. MethylBERT finds the optimal transformation of local estimates to reduce the skewness, whereas CancerDetector iteratively removes regions that are outside of the standard deviation. Therefore, we have specifically compared the MethylBERT estimation adjustment and CancerDetector removal of confounding factors using the same $P(cell\ type|read)$ values (Figure 4D). MethylBERT estimation adjustment outperformed CancerDetector removal of confounding factors in terms of median absolute error between the ground truth and estimated purities. Although CancerDetector removal of confounding factors cannot improve the estimated values when the tumour is the major cell type, MethylBERT is able to make a better adjustment regardless of ground-truth tumour purity. When the adjustment is conducted, MethylBERT calculates the Fisher information in each region because the maximum log-likelihood estimation is applied locally rather than globally. For the 20 pseudo-bulks, the median of the Fisher information still correlates with the ground-truth tumour purity after the estimation is adjusted (Figure 4E).

In addition to the accurate inference of mixture proportions, MethylBERT can reconstruct the sample-specific tumour and normal cell-type methylation patterns, in both bulk and read-level. For example, we compared the MethylBERT-estimated tumour and normal cell-type methylation beta-values for a pseudo-bulk sample to the reference beta-values obtained from the training data (Figure 4F). At the level of single CpGs within one region, MethylBERT successfully reconstructed the methylation beta-values of the reference profile, although the methylation patterns in chr9:139740205-139740700 show a high complexity in the normal cell type. The estimated beta values were also more similar to the correct reference cell-type profile according to the hierarchical clustering result (Supplementary Figure 5A). Moreover, we found that MethylBERT successfully dissected tumour and normal methylation patterns over DMRs in all pseudo-bulk samples (Supplementary Figure 5B and D). In order to make sure that the result is not influenced by dominant tumour hypermethylation over the selected DMRs, we tested it again with 100 regions where half is tumour hypermethylated and the other half is hypomethylated regions. MethylBERT still successfully dissected cell type-specific methylation patterns in these tests (Supplementary Figure 5C and E).

MethylBERT enables accurate circulating tumour DNA analysis by detecting rare tumour signals

Besides bulk tumour deconvolution, precise sequencing-based tumour deconvolution is required in ctDNA analysis due to the extremely low quantity of tumour-derived DNA found in liquid biopsies at early disease stages. Hence, we evaluated MethylBERT as an early cancer detection algorithm for blood plasma samples and compared its performance to other previous methods.

To validate whether MethylBERT can detect a very low percentage of tumour-specific signals, we further generated 10 pseudo-bulks by mixing reads from non-neoplastic B cell and DLBCL samples with a tumour ratio smaller than 1%. MethylBERT achieved a lower median absolute error than the other methods (Figure 5A). All methods can estimate tumour purities proportional to the ground-truth values with the exception of DISMIR applied with its own selected regions but MethylBERT shows the highest correlation between two values with p-value < 0.01 (Figure 5B). Therefore, we demonstrated that MethylBERT is sufficiently sensitive for ctDNA analysis compared to the currently available methods.

As an application to real ctDNA samples, we used targeted BS-seq data collected from 14 healthy donors and 40 colorectal cancer (CRC) patients in five different stages (GSE149438) (Supplementary Figure 6A). The estimated tumour purity significantly differs between healthy donors and tumour patients later than stage Ⅰ (Figure 5C). This implies that MethylBERT can be instrumental for ctDNA tumour diagnosis in some early stages (Ⅱ-Ⅲ) of CRC patients. The median value of estimated tumour purities has an explicit discrepancy between the healthy donors and all stages of CRC patients.

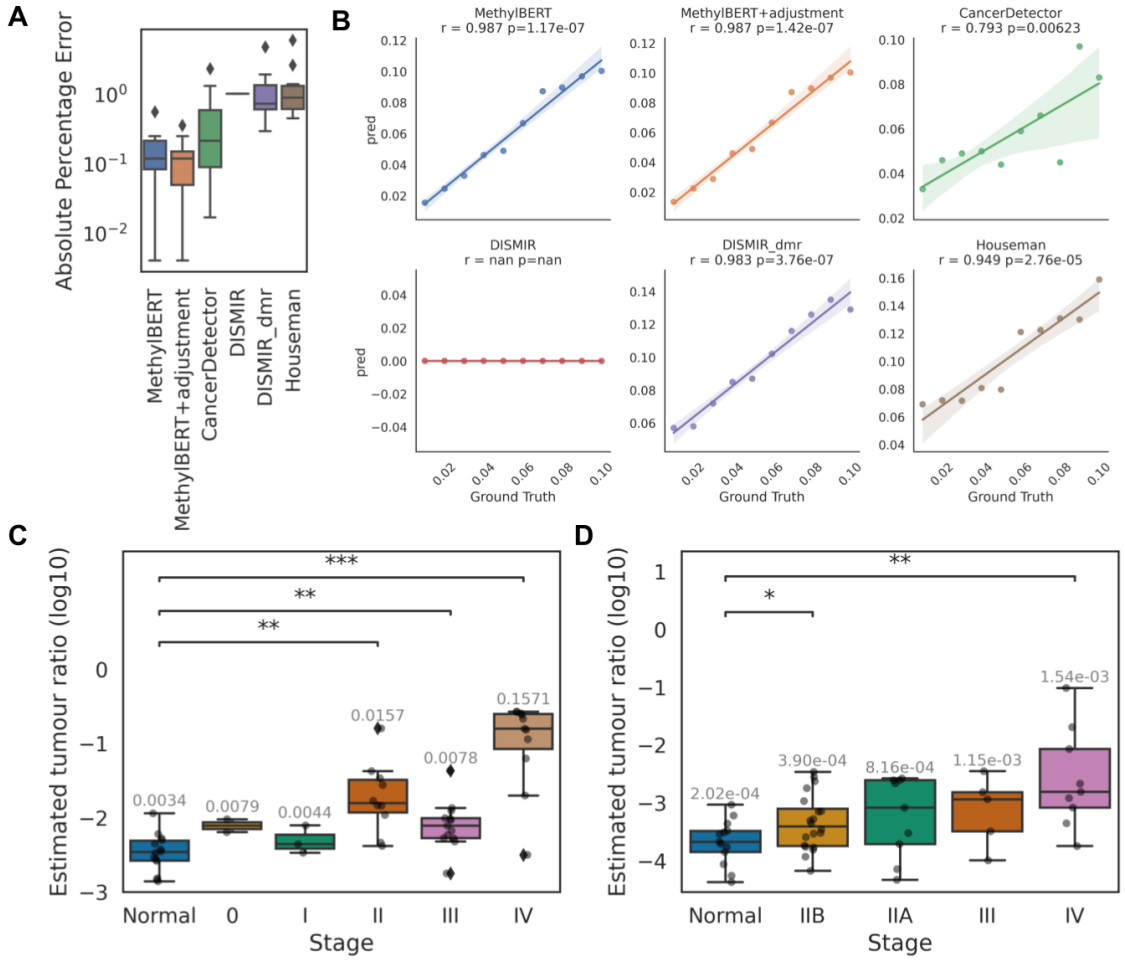

Figure 5. ctDNA tumour deconvolution results. ***A.*** *Deconvolution of simulated pseudo-bulk samples with a low percentage of tumour DNA. Distribution of absolute percentage error values in each method.* ***B.*** *Comparison of ground-truth and estimated tumour purities in each method. For each comparison, Spearman's correlation is given with a respective p-value.* ***C-D.*** *Tumour deconvolution results in ctDNA samples from CRC patients (C) and PDAC patients (D). Both analyses include healthy donors as well. The median of estimated tumour purities in each stage is written at the top of the box plot. Two-sided Mann-Whitney-Wilxcoxon test p-values between each stage and healthy donors are denoted with stars. '*', '**', and '***' mean p-value ≤ 0.05, ≤ 0.01, and ≤ 0.001, respectively.*

From the same dataset, we also collected 44 pancreatic ductal adenocarcinoma (PDAC) patients in four different stages (ⅡA, ⅡB, Ⅲ and Ⅳ) and conducted the same analysis using MethylBERT (Supplementary Figure 6B and Figure 5D). PDAC is widely recognised as one of the trickiest cancer types to be identified during the early stages, which is also confirmed by the original analysis in this dataset [26]. The median tumour content value estimated by MethylBERT is higher in every stage of PDAC cohorts than in healthy donors. Among the early-stage patients, stage ⅡB showed a statistically significant difference from the healthy donors. Overall, in both CRC and PDAC ctDNA analyses, all early-stage cancer patient samples except for CRC stage Ⅱ have a median value lower than 0.01 demonstrating the necessity of sensitive sequencing-based tumour deconvolution models in ctDNA methylation analysis.

Discussion

DNA methylation undergoes profound changes during tumorigenesis [1,19] resulting in highly specific methylation patterns in tumour cells. Sequencing data is particularly valuable in this respect, since it can deliver DNAm patterns at single-molecule resolution. The broad genomic coverage and the preservation of rare cell-type signals extend the usage of read-level DNAm data for various applications in cancer research, including deconvolution of bulk tumour methylomes and circulating tumour DNA analysis using liquid biopsy.

Aiming to fully utilise the potential of sequencing-based data, we developed MethylBERT, a new Transformer-based model for read-level tumour methylation pattern identification. Based on the estimated posterior probability by the MethylBERT model, the tumour purity of each bulk sample can be inferred by maximum likelihood estimation. The estimated purity can be adjusted for a more accurate inference by considering region-wise tumour purity estimation.

In the benchmarking using simulated read-level methylation patterns with different scenarios, MethylBERT achieved the best performance regardless of methylation pattern complexity, read length and read coverage. When it comes to pseudo-bulk deconvolution, MethylBERT also achieved the lowest absolute error among all evaluated methods. Furthermore, the tumour deconvolution results for very low tumour purity (<1%) confirm that MethylBERT is capable of detecting rare tumour signals and is applicable to ctDNA analysis. With real blood plasma samples collected from healthy donors and cancer patients, MethylBERT could distinguish some early-stage patients from healthy donors. This assures the applicability of MethylBERT as an early tumour diagnosis model. Our results demonstrate not only the remarkable performance of MethylBERT in varying experiments but also provide in-depth analyses of BERT pre-training and the dynamics of the estimated posterior probability during model training. The findings enhance the degree of understanding about the MethylBERT model and explain why the model successfully detects tumour-specific methylation patterns. To sum up, MethylBERT is the first Transformer-based model for encoding and classification of read-level methylation patterns. We are convinced that it will become a valuable tool in the field of cancer research, extending the usability of various types of DNA methylation sequencing data. MethylBERT is agnostic with respect to the source of read-level methylation data, and can be applied to both solid tumour tissues and blood plasma samples as well as non-cancerous bulk methylomes, unlike previous methods that focused on only specific sample types [12,13,20].

In the future, different extensions of MethylBERT are possible. The current version only infers the tumour purity within bulk samples, but, with further modifications of the model, deconvolution of multiple or unknown cell types within a bulk could be accomplished. Cell-type composition of the tumour microenvironment (TME) consisting of cancer, immune and stromal cells, is crucial for defining clinically relevant tumour subtypes or drug development in immunotherapy [27,28]. DNAm data-based TME deconvolution has been done by using array-based data [29,30] and provided valuable insights in various cancer studies [31,32]. However, sequencing-based TME methylation deconvolution still has not been extensively applied. We expect that the advanced version of MethylBERT handling

multiple cell types can be applied to TME deconvolution and cover a broader range of tumour-related deconvolution analyses.

Furthermore, similarly to other Transformer-based models, MethylBERT features a large number of model parameters, which creates a considerable computational overhead during the training. We are currently working on a computationally less demanding version of the MethylBERT model with a reduced number of parameters. This work is particularly important for long-read sequencing data analysis. Long-read sequencing broadens our perspectives on genome and sequential methylation patterns by generating reads of several thousands to millions of base pairs [33,34]. Based on the results of read-level methylation classification with 500bp-long simulated reads, we believe that MethylBERT can handle differentially methylated patterns being a part of a read-level sequence, and will be able to provide accurate and robust analyses of tumour-specific methylation patterns in long-read sequencing data. However, the BERT model demands excessive computational time and memory to process long sequences. This issue has been extensively discussed, and several methods have been proposed to reduce the computational cost of BERT [35–37]. Therefore, in the future, we plan to apply diverse types of advanced BERT models with fewer parameters to analyse long-read sequencing data.

Methods

MethylBERT model

BERT model

The basis of MethylBERT is the Bidirectional Encoder Representations from Transformers (BERT) model which has achieved groundbreaking progress in various natural language processing studies [38,39]. Transformers [40] in the BERT model take the attention mechanism resembling human cognitive attention and map given query (Q), key (K) and value (V) sequences to an output sequence. Transformers particularly use scaled dot-product attention [41], where the dot-product of query and key sequences are scaled by the inverse square root of the dimension of key vectors:

$$Attention(Q,\ K,\ V)\ =\ softmax(\frac{QK^T}{\sqrt{d_k}})V.$$

The scale factor enables the model to avoid an extremely large magnitude of multiplied value which especially occurs with a long sequence input. Multi-head attention which concatenates H multiple attentions calculated from the weighted query, key and value sequences is a major advantage of Transformers in learning different projections of these three sequences:

$$A_i\ =\ Attention_i(QW^Q{}_i,\ KW^K{}_i,\ VW^V{}_i),$$

$$Multi-head\ attention(Q,\ K,\ V)\ =\ Concatenation(A_1,\ldots,A_H)W^O$$

where A_i refers to an attention layer. In every attention layer, given query, key and value matrices are projected using the parameter matrices $W^Q{}_i$, $W^K{}_i$ and $W^V{}_i$. Then, multi-head

attention is created as a concatenation of all H attention layers projected using another parameter matrix W^O.

We modified the BERT model to process DNA sequence fragments and sequential methylation patterns (Supplementary Figure 1). Originally, BERT needs three types of input embeddings: token embeddings, segment embeddings and position embeddings. In MethylBERT, we replaced the token embeddings assigned for tokenised words with DNA token embeddings. The segment embeddings indicate the sentence label of each token and are used for NSP. Instead, for MethylBERT, we created methylation embeddings for encoded methylation patterns. The position embeddings were also used in MethylBERT to guide the position of tokens in the sequencing read.

Pre-training MethylBERT

As described above, BERT requires long pre-training to learn the general context of input data and avoids heavily engineered task-specific architectures for every specific task. The pre-training step in MethylBERT is inspired by DNABERT [16]. We split the hg19 genome into 510bp segments and generated a 3-mers sequence of each segment, and the 3-mers segment is referred to as a token. While Ji et al. randomly sampled the read length between 5 and 510 with a certain probability for DNABERT [16], we used a fixed value of the sequence length 510 since it did not make a major performance change. For pre-training, we only applied a masked language model (MLM) by randomly masking 15% of 3-mers. Three masking schemes were employed following the original BERT paper [14]: 80% of selected 3-mers were masked with [MASK] token, 10% were replaced with another randomly chosen token, and the rest was unchanged. We masked left and right tokens together with the selected token in order to prevent a biased model predicting the masked token from neighbouring tokens. 3-mers tokens have 69 labels in total including five special tokens. A categorical cross-entropy loss L was calculated over all tokens $t \in \{1 \ldots T\}$ for every step of the model optimisation:

$$L_{pre-training} = -\sum_{t=1}^{T}\sum_{l=1}^{69} m^t y^t_l \, log(\hat{y}^t_l)$$

where y^t_l and $\hat{y}^t_l$ refer to as one-hot encoded value of label l in token t and corresponding logit value calculated by the MethylBERT model. m^t is a binary value indicating whether token t is masked or not.

We pre-trained the MethylBERT model for 120k steps including 10k warm-up steps and 20k decrease steps at the end of the training with a learning rate of 4e-4. The batch size was set to 256 and the gradient was accumulated over 4 steps. In the network architecture, the hidden layer had a size of 768 and the baseline model had 12 layers with 12 attention heads. However, since the smaller size BERT models performed similarly to the baseline model in the read classification analysis using simulated data, we used 6 layers of the MethylBERT model for 500 bps read analysis (Supplementary Figure 7). AdamW optimiser [42] was used with decay rate 0.01, β_1=0.9 and β_2=0.98 values. Since reference genomes do not have methylation patterns, we filled the methylation embeddings up with zeros during pre-training.

Fine-tuning for read-level methylation pattern classification

For fine-tuning, DNA sequence fragments from reads are processed into 3-mers sequences as described in the previous section. The input length was reduced to 150 due to the shorter sequencing read length. In order to represent CpG methylomes, three numbers were used for methylation pattern encoding: 0 for unmethylated CpGs, 1 for methylated CpGs, and 2 for non-CpGs. CpG methylation patterns were assigned to the 3-mers where the cytosine of CpG is located in the middle.

The encoder part of the MethylBERT network takes DNA, methylation and position embeddings, and generates an encoded vector of $sequence_length \times 768$ dimension. This vector is concatenated with DMR information embedded into a space of dimension $sequence_length$, and the following the cell-type classifier calculates the posterior probabilities of cell types using the concatenated vectors. The cell type with the highest posterior probability is determined as the classification outcome.

The MethylBERT model was fine-tuned over 600 or 1000 steps with a learning rate of 4e-4. We used the same optimisation scheme as for pre-training but applied the cross-entropy loss on the cell-type label for read-level methylome classification:

$$L_{fine-tuning} = -\sum_{r=1}^{R} \sum_{c \in \{T, N\}} m_{r,c} log \frac{exp(x^{c}_{r})}{\sum_{c' \in \{T, N\}} exp(x^{c'}_{r})}$$

where x^{c}_{r} is the final activation for the cell type c (T=tumour, N=normal) for each read r. m^{c}_{r} is a binary value from one-hot encoded cell-type label c. The logits are normalised with a softmax function before the cross-entropy loss is calculated.

Tumour purity/fraction estimation

Since the proportions of the tumour and non-tumour compartments sum up to one, we used a single-parameter likelihood function to estimate the best tumour purity $\hat{\delta}$ from collected reads $\{r_1, ..., r_N\}$:

$$L(\delta) = \prod_{i=1}^{N} [\delta P(r_i | cell\ type = Tumour) + (1 - \delta) P(r_i | cell\ type = Normal)],$$

$$\hat{\delta} = argmax_{\delta} L(\delta).$$

The MethylBERT model calculates only the posterior probabilities of cell types given a read, so Bayes' theorem is applied to calculate the posterior probability of a read assuming that every read has the same marginal probability:

$$P(r_i | cell\ type = T) \propto P(cell\ type = T | r_i) P(cell\ type = T)^{-1}.$$

The prior probability of the cell types is calculated from the training dataset used for fine-tuning. To avoid the high complexity of likelihood computation and the Fisher information calculation, we use the log-likelihood function for maximum likelihood estimation. We employ a grid-search algorithm to find the optimal parameter $\hat{\delta}$ increasing the δ value by 0.0001 from zero to one.

As shown in Supplementary Figure 8A, the ratio of tumour-derived reads in DMRs does not have a bell-shaped distribution when tumour-normal cell types do not have an equal proportion. However, estimating the tumour purity only using the log-likelihood function above assumes that tumour-derived reads are equally distributed in every DMR. Therefore, we propose an adjustment of estimated tumour purity by finding a transformation $\widehat{W}$ that maps the estimated tumour purity in K regions, $\overline{\delta} = \{\delta_1, \ldots, \delta_K\}$, into a distribution with a minimum skewness. The idea stems from the strong correlation between the ground-truth tumour purity and the skewness of region-wise tumour purity (Supplementary Figure 8B). Assuming that the DMR k includes N reads, $\{r^k_1, \ldots, r^k_N\}$, the optimal transformation $\widehat{W}$ is found as follows:

$$\delta_k = argmax_{\delta} \prod_{r^k_i \in \{r^k_1, \ldots, r^k_N\}} [\delta P(r^k_i | cell\ type = Tumour) + (1 - \delta) P(r^k_i | cell\ type = Normal),$$

$$\widehat{W} = argmin_W\, \mu_3(W\overline{\delta}),$$

$$\mu_3(X) = \frac{E[X-\mu]^3}{(E[X-\mu]^2)^{3/2}},\ \mu = E(X)$$

where $E(x)$ refers to as the expectation of given random variable x. The expectation-maximisation (EM) algorithm is used to find the optimal transformation $\widehat{W}$. Once the best transformation is found, we assume that the distribution of local estimates is bell-shaped and determine the final estimation of tumour purity as the mean value of transformed region-wise tumour purities:

$$\widehat{\delta} = \frac{1}{K}\widehat{W}^T\overline{\delta}.$$

The Fisher information indicates the amount of information about a model parameter carried by observed variables. In the case of a one-dimensional model parameter, the Fisher information is equivalent to an estimate of the model precision [25]. It is calculated as the variance of the derivative of the log-likelihood function with respect to the model parameter:

$$FI(\delta) = Var[\frac{\partial}{\partial\delta} logL(\delta)].$$

When the tumour purity adjustment is applied, the Fisher information cannot be calculated for the final estimation but in each region. Therefore, in this case, MethylBERT provides the Fisher information values as many as the number of selected DMRs.

Data preparation

Diffuse large B cell lymphoma WGBS data

Diffuse large B cell lymphoma (DLBCL) and non-neoplastic B-cell bisulfite sequencing data were downloaded from the Gene Expression Omnibus (GEO) database with the accession number GSE137880 [43]. 8 samples from two DLBCL patients and 8 non-neoplastic B cell samples from two donors were used for the experiments. From both cell types, 4 samples were assigned for the training and validation dataset and the rest were used for creating pseudo-bulks. All downloaded FastQ files were aligned with hg19 reference genome by

Bismark 0.22.3 [44] after trimming using Trim Galore 0.6.6 [45], then duplicated reads were removed using picard Mark Duplicates 1.141. We followed the whole pipeline specified in our previous benchmarking study [10].

CRC and PDAC BS-seq data

To train each MethylBERT model for ctDNA analysis derived from CRC and PDAC patients, we downloaded single-cell BS-seq of CRC cells with the GEO accession number GSE97693 [46] and WGBS of PDAC tissue with GSE63123 [47]. Only part of the CRC samples (341 cells) were downloaded as a training dataset. Similarly, we utilised only a subset of the PDAC data set involving seven samples collected from primary tumours in three different patients. During the data processing, we used TrimGalore 0.6.6 for adaptor trimming and quality control. Afterwards, the samples were aligned with the hg19 reference genome using Bismark 0.22.3. Duplicate removal was done by picard Mark Duplicates 1.141.

CRC, PDAC and control healthy ctDNA blood plasma data

For ctDNA experiments, we downloaded targeted BS-seq of plasma samples from gastrointestinal cancer patients and controlled healthy donors with the GEO accession number GSE149438 [26]. The downloaded dataset includes samples from 46 healthy donors, 74 PDAC patients, and 40 CRC patients in different cancer stages. We only used 44 PDAC patients whose cancer stage is clarified for the experiment. 32 healthy plasma samples were used to fine-tune the MethylBERT model and the rest was used as a comparison group in the tumour diagnosis analyses. We processed the data via trimming using TrimGalore, alignment to hg19 using Bismark and duplicate removal using picard Mark Duplicates. The samples were aligned with paired-end mode and the unmapped reads were re-aligned with single-end mode.

DMR calling

Selecting informative regions with tumour-specific signals is vital in tumour purity estimation. For MethylBERT, we pre-selected differentially methylated regions (DMRs) to collect reads presenting informative methylation patterns. Tumour-specific DMRs were called by comparing tumour samples to non-tumour samples and the DSS package was used for the calling [48]. Parameters were set up as follows: delta value 0.2, P-value threshold 0.05, minimum number of CpGs 4, minimum length 50 bps and distance to merge 50 bps. In all analyses, we picked the top 100 DMRs based on the highest areaStat score. DSS performs a Wald test to identify differentially methylated loci and the areaStat score is calculated by summing all the test statistics up within each DMR. Therefore, a higher value of areaStat is likely to secure a larger number of CpGs in the region that are significantly differentially methylated between tumour and normal cell types.

Read-level methylome simulation

Simulated read-level methylomes were used to mimic different scenarios of tumour-specific signals for the evaluation of methylation pattern classification. We selected 100 CpG islands with the highest number of CpGs as the regions where reads are sampled. For each region, two mean methylation levels need to be assigned for tumour and normal cell types. We sampled a tumour mean methylation value d_i from a beta distribution whose β parameter is fixed to 5 and assigned $1 - d_i$ to a normal mean methylation value. Then, read-level methylation patterns were sampled from a binomial distribution with the probability $1 - d_i$ and d_i for normal and tumour cell types, respectively. The entire mechanism of sampling a read-level methylome with *K* CpGs, $\overline{m}^T = \{m^T_1, ..., m^T_K\}$ for tumour cell type and $\overline{m}^N = \{m^N_1, ..., m^N_K\}$ for normal cell type, is described as follows:

$$d_i \sim Beta(\alpha, \beta = 5),$$

$$\overline{m}^T \sim Binomial(n = K, p = d_i),$$

$$\overline{m}^N \sim Binomial(n = K, p = 1 - d_i)$$

where α and β are two shape parameters of the beta distribution, and n and p are the number of trials and the success probability of one trial in the binomial distribution.

Four different α values, 0.1, 1.0, 2.0 and 3.0, were used in the beta distribution to model different complexities of methylation patterns. Since a larger alpha value increases the variance of the beta distribution with a fixed beta value, it is more likely that a lower tumour methylation value d_i will be sampled. This makes a higher methylation value $1 - d_i$ assigned to normal cell type. The smaller gap between d_i and $1 - d_i$ forms more complex methylation patterns between tumour and normal cell types (Supplementary Figure 2).

When CpG-specific methylation patterns are sampled (Supplementary Figure 2D), d_i and $1 - d_i$ are assigned to the odd and even indices of CpGs. Therefore, in this simulation, the average methylation level between tumour and normal cell types do not significantly differ, but still have cell type-specific methylation patterns.

Evaluation of read classification performance

The read classification performance evaluation includes an HMM-based algorithm. We designed the HMM to take methylation patterns as an observation, thus the observation has two categories: methylated and unmethylated CpGs. The hidden state is also a two-categorical variable with the assumption of whether the CpG is differentially methylated or not between tumour and normal cell types.

Although CancerDetector [20] and DISMIR [13] do not explicitly mention 'read classification', CancerDetector calculates $P(read|cell\ type\ =\ Tumour)$ whereas DISMIR has a 'd-score' that quantifies the chance that a given read is derived from tumour cells as interpreted by the authors. Therefore, we conducted read classification for those methods by assigning every read to the tumour class when the probability or d-score > 0.5.

Code availability

The read-level methylation simulation code is available at https://github.com/CompEpigen/methylseq_simulation. The MethylBERT code will be uploaded to GitHub in the near future.

References

1. *Guo M, Peng Y, Gao A, Du C, Herman JG. Epigenetic heterogeneity in cancer. Biomark Res. 2019;7: 23. doi:10.1186/s40364-019-0174-y*
2. *Zhao L, Wu X, Zheng J, Dong D. DNA methylome profiling of circulating tumor cells in lung cancer at single base-pair resolution. Oncogene. 2021;40: 1884–1895. doi:10.1038/s41388-021-01657-0*
3. *Meyer B, Clifton S, Locke W, Luu P-L, Du Q, Lam D, et al. Identification of DNA methylation biomarkers with potential to predict response to neoadjuvant chemotherapy in triple-negative breast cancer. Clin Epigenetics. 2021;13: 226. doi:10.1186/s13148-021-01210-6*
4. *Zhang W, Feng H, Wu H, Zheng X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics. 2017;33: 2651–2657. doi:10.1093/bioinformatics/btx303*
5. *Fiala C, Diamandis EP. Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. BMC Med. 2018;16: 166. doi:10.1186/s12916-018-1157-9*
6. *Shu C, Zhang X, Aouizerat BE, Xu K. Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. Epigenetics Chromatin. 2020;13: 51. doi:10.1186/s13072-020-00372-6*
7. *Yan Y-Y, Guo Q-R, Wang F-H, Adhikari R, Zhu Z-Y, Zhang H-Y, et al. Cell-Free DNA: Hope and Potential Application in Cancer. Front Cell Dev Biol. 2021;9: 639233. doi:10.3389/fcell.2021.639233*
8. *Duffy MJ, Crown J. Circulating Tumor DNA as a Biomarker for Monitoring Patients with Solid Cancers: Comparison with Standard Protein Biomarkers. Clin Chem. 2022;68: 1381–1390. doi:10.1093/clinchem/hvac121*
9. *Yang Y, Zhang T, Wang J, Wang J, Xu Y, Zhao X, et al. The clinical utility of dynamic ctDNA monitoring in inoperable localized NSCLC patients. Mol Cancer. 2022;21: 117. doi:10.1186/s12943-022-01590-0*
10. *Jeong Y, de Andrade E Sousa LB, Thalmeier D, Toth R, Ganslmeier M, Breuer K, et al. Systematic evaluation of cell-type deconvolution pipelines for sequencing-based bulk DNA methylomes. Brief Bioinformatics. 2022;23. doi:10.1093/bib/bbac248*

11. *Lee D, Lee S, Kim S. PRISM: methylation pattern-based, reference-free inference of subclonal makeup. Bioinformatics. 2019;35: i520–i529. doi:10.1093/bioinformatics/btz327*

12. *Zheng X, Zhao Q, Wu H-J, Li W, Wang H, Meyer CA, et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biol. 2014;15: 419. doi:10.1186/s13059-014-0419-x*

13. *Li J, Wei L, Zhang X, Zhang W, Wang H, Zhong B, et al. DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads. Brief Bioinformatics. 2021;22. doi:10.1093/bib/bbab250*

14. *Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. doi:10.48550/arxiv.1810.04805*

15. *Shi W, Demberg V. Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. pp. 5789–5795. doi:10.18653/v1/D19-1586*

16. *Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37: 2112–2120. doi:10.1093/bioinformatics/btab083*

17. *Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18: 67. doi:10.1186/s13059-017-1189-z*

18. *Wang Z, Yin J, Zhou W, Bai J, Xie Y, Xu K, et al. Complex impact of DNA methylation on transcriptional dysregulation across 22 human cancer types. Nucleic Acids Res. 2020;48: 2287–2302. doi:10.1093/nar/gkaa041*

19. *Liu Y, Gu Y, Su M, Liu H, Zhang S, Zhang Y. An analysis about heterogeneity among cancers based on the DNA methylation patterns. BMC Cancer. 2019;19: 1259. doi:10.1186/s12885-019-6455-x*

20. *Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res. 2018;46: e89. doi:10.1093/nar/gky423*

21. *Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, et al. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics. 2020;112: 144–150. doi:10.1016/j.ygeno.2019.05.007*

22. *Paska AV, Hudler P. Aberrant methylation patterns in cancer: a clinical view. Biochem Med (Zagreb). 2015;25: 161–176. doi:10.11613/BM.2015.017*

23. *Clark K, Khandelwal U, Levy O, Manning CD. What Does BERT Look At? An Analysis of BERT's Attention. arXiv. 2019. doi:10.48550/arxiv.1906.04341*

24. *Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13: 86. doi:10.1186/1471-2105-13-86*

25. *Fujita K, Okada K, Katahira K. The Fisher information matrix: A tutorial for calculation for decision making models. 2022. doi:10.31234/osf.io/hdwut*

26. *Kandimalla R, Xu J, Link A, Matsuyama T, Yamamura K, Parker MI, et al. EpiPanGI Dx: A Cell-free DNA Methylation Fingerprint for the Early Detection of Gastrointestinal Cancers. Clin Cancer Res. 2021;27: 6135–6144. doi:10.1158/1078-0432.CCR-21-1982*

27. *Singh O, Pratt D, Aldape K. Immune cell deconvolution of bulk DNA methylation data reveals an*

association with methylation class, key somatic alterations, and cell state in glial/glioneuronal tumors. Acta Neuropathol Commun. 2021;9: 148. doi:10.1186/s40478-021-01249-9

28. *Tang T, Huang X, Zhang G, Hong Z, Bai X, Liang T. Advantages of targeting the tumor immune microenvironment over blocking immune checkpoint in cancer immunotherapy. Signal Transduct Target Ther. 2021;6: 72. doi:10.1038/s41392-020-00449-4*
29. *Zhang Z, Christensen BC, Salas LA. Abstract 1212: ExTIME: Extended tumor immune micro-environment cell mixture deconvolution using DNA methylation and a novel tumor-site-specific hierarchical approach. Cancer Res. 2022;82: 1212–1212. doi:10.1158/1538-7445.AM2022-1212*
30. *Zhang Z, Wiencke JK, Kelsey KT, Koestler DC, Christensen BC, Salas LA. HiTIMED: hierarchical tumor immune microenvironment epigenetic deconvolution for accurate cell type resolution in the tumor microenvironment using tumor-type-specific DNA methylation data. J Transl Med. 2022;20: 516. doi:10.1186/s12967-022-03736-6*
31. *Hey J, Halperin C, Hartmann M, Mayer S, Schönung M, Lipka DB, et al. DNA methylation landscape of tumor-associated macrophages reveals pathways, transcription factors and prognostic value relevant to triple-negative breast cancer patients. Int J Cancer. 2023;152: 1226–1242. doi:10.1002/ijc.34364*
32. *Vougiouklakis T, Aung PP, Vasudevaraja V, Prieto VG, Torres-Cabala CA, Sulman EP, et al. Correlative study of epigenetic regulation of tumor microenvironment in spindle cell melanomas and cutaneous malignant peripheral nerve sheath tumors. Sci Rep. 2020;10: 12996. doi:10.1038/s41598-020-69787-1*
33. *Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21: 597–614. doi:10.1038/s41576-020-0236-x*
34. *Gouil Q, Keniry A. Latest techniques to study DNA methylation. Essays Biochem. 2019;63: 639–648. doi:10.1042/EBC20190027*
35. *Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, et al. Big Bird: Transformers for Longer Sequences. Advances in Neural Information Processing Systems. 2020.*
36. *Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. 2019. doi:10.48550/arxiv.1910.01108*
37. *Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. arXiv. 2020. doi:10.48550/arxiv.2004.05150*
38. *Aum S, Choe S. srBERT: automatic article classification model for systematic review using BERT. Syst Rev. 2021;10: 285. doi:10.1186/s13643-021-01763-w*
39. *Koroteev MV. BERT: A Review of Applications in Natural Language Processing and Understanding. arXiv. 2021. doi:10.48550/arxiv.2103.11943*
40. *Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017. doi:10.48550/arxiv.1706.03762*
41. *Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. arXiv. 2015. doi:10.48550/arxiv.1508.04025*
42. *Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. arXiv. 2017. doi:10.48550/arxiv.1711.05101*
43. *Do C, Dumont ELP, Salas M, Castano A, Mujahed H, Maldonado L, et al. Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. Genome Biol. 2020;21: 153. doi:10.1186/s13059-020-02059-3*

44. *Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27: 1571–1572. doi:10.1093/bioinformatics/btr167*

45. *Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, et al. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo. 2023. doi:10.5281/zenodo.7598955*

46. *Adossa N, Khan S, Rytkönen KT, Elo LL. Computational strategies for single-cell multi-omics integration. Comput Struct Biotechnol J. 2021;19: 2588–2596. doi:10.1016/j.csbj.2021.04.060*

47. *McDonald OG, Li X, Saunders T, Tryggvadottir R, Mentch SJ, Warmoes MO, et al. Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. Nat Genet. 2017;49: 367–376. doi:10.1038/ng.3753*

48. *Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32: 1446–1453. doi:10.1093/bioinformatics/btw026*