*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

# Identifying differentially methylated sites in samples with varying tumor purity

**Antti Häkkinen [1,*], Amjad Alkodsi [1], Chiara Facciotto [1], Kaiyang Zhang [1], Katja Kaipio [2], Sirpa Leppä [3], Olli Carpén [4,5,6], Seija Grénman [7], Johanna Hynninen [7], Sakari Hietanen [7], Rainer Lehtonen [1], and Sampsa Hautaniemi [1,*]**

[1] Research Programs Unit, Genome-Scale Biology, Medicum and Department of Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, Helsinki, Finland [2] Department of Pathology and Forensic medicine, Institute of Biomedicine, University of Turku, Turku, Finland [3] Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland [4] Department of Pathology, Medicum, University of Helsinki and HUSLAB, Helsinki University Hospital, Helsinki, Finland [5] Research Programs Unit, Genome-Scale Biology, University of Helsinki, Helsinki, Finland [6] Department of Pathology, Institute of Biomedicine, University of Turku and Turku University Hospital, Turku, Finland [7] Department of Obstetrics and Gynecology, University of Turku and Turku University Hospital, Turku, Finland.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** DNA methylation aberrations are common in many cancer types. A major challenge hindering comparison of patient-derived samples is that they comprise of heterogeneous collection of cancer and microenvironment cells. We present a computational method that allows comparing cancer methylomes in two or more heterogeneous tumor samples featuring differing, unknown fraction of cancer cells. The method is unique in that it allows comparison also in the absence of normal cell control samples and without prior tumor purity estimates, as these are often unavailable or unreliable in clinical samples.
**Results:** We use simulations and next-generation methylome, RNA, and whole-genome sequencing data from two cancer types to demonstrate that the method is accurate and outperforms alternatives. The results show that our method adapts well to various cancer types and to a wide range of tumor content, and works robustly without a control or with controls derived from various sources.
**Availability:** The method is freely available at https://bitbucket.org/anthakki/dmml.
**Contact:** antti.e.hakkinen@helsinki.fi, sampsa.hautaniemi@helsinki.fi
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Aberrant DNA methylation is a hallmark of all cancer types (Hanahan and Weinberg, 2011; Witte *et al.*, 2014; Shen and Laird, 2013; Timp and Feinberg, 2013). Compared to genomic alterations, DNA methylation offers a more flexible yet a persistent mechanism to exert changes on the phenotype, which manifests in silencing tumor suppressor genes, activating proto-oncogenes, or causing chromosomal instability (Witte *et al.*, 2014; Shen and Laird, 2013; Timp and Feinberg, 2013; Yang *et al.*, 2015). While some patterns have been identified, the role of DNA methylation alterations in cancer development, tumor pathogenesis, and treatment response varies between cancer types (Witte *et al.*, 2014; Yang *et al.*, 2015; Ciriello *et al.*, 2013). Advances in next-generation sequencing (NGS) in combination with classical bisulfite conversion (Frommer *et al.*, 1992; Harris *et al.*, 2010) have allowed profiling methylomes at a single nucleotide resolution at an unprecedented scale (Shen and Laird, 2013; Ciriello *et al.*, 2013). These developments have surged an interest to develop personalized clinical applications that employ DNA methylation alterations as diagnostic and prognostic biomarkers and as therapeutic targets (Wei *et al.*, 2006; Altman *et al.*, 2013). A major challenge in this is that the surgically removed samples comprise of heterogeneous mixture of cancer cells and the microenvironment. As the exact tumor content (tumor cell fraction, tumor purity) and cell composition varies considerably

 

between the samples, direct comparison of samples even from the same patient without correcting for the tumor cell content can lead to spurious results (Carter *et al.*, 2012; Aran *et al.*, 2015; Zheng *et al.*, 2017).

Immunohistochemical staining has been the most used technique to determine cell composition in tissue sections and to select high purity samples for further experiments. More recently, high-throughput measurement technologies have enable cost-efficient and rapid production of patient-derived molecular data, but also allow estimating the sample tumor content. Computational tools have been developed for purity estimation using somatic copy-number data (Carter *et al.*, 2012; Van Loo *et al.*, 2010), single-nucleotide polymorphisms (Van Loo *et al.*, 2010), variant allele frequency of somatic mutations (Carter *et al.*, 2012), or RNA expression (Yoshihara *et al.*, 2013). Evidence suggests that these methods outperform manual analysis and allow analysis of large-scale datasets (Carter *et al.*, 2012; Zheng *et al.*, 2017). While the tools allow fast and reproducible tumor purity analysis, most differential DNA methylation analysis methods i) do not account for the sample heterogeneity or adjust the analysis for low tumor purity or differences between samples (Hebestreit *et al.*, 2013; Hansen *et al.*, 2012; Feng *et al.*, 2014; Sun *et al.*, 2014; Sun and Yu, 2016; Wang *et al.*, 2016), ii) require a library of control samples (Houseman *et al.*, 2012), or iii) do not account for co-methylation of closely located sites, cannot correct for this at single-nucleotide level (Hebestreit *et al.*, 2013; Hansen *et al.*, 2012; Feng *et al.*, 2014), or model co-methylation uniformly (Sun and Yu, 2016; Wang *et al.*, 2016; Zheng *et al.*, 2014). All the shortcomings lead to biased findings and false biological interpretation. The requirement for controls is problematic, as in many cases appropriate controls are unavailable or incomparable, and predictions using universal normal controls poorly correlate with those of matched controls in many cancer types (Zheng *et al.*, 2017).

Here, we present a method based on a latent stochastic model which allows comparing DNA methylomes at a single-nucleotide resolution between two or more tumor samples with different, unknown tumor purities. The method performs accurately without a normal cell control sample or prior tumor purity estimates, and accounts for spatially co-methylated cytosines, improving accuracy at lower coverage sites. We demonstrate the superior performance using simulations and two sets of NGS cancer data — targeted sequencing data from high-grade serous ovarian cancer patients and genome-wide reduced representation bisulfite sequencing data from diffuse large B-cell lymphoma patients.

## 2 Models and methods

Patient-derived samples are composed of different cell types, such as various stromal and immune cell types in addition to cancer cells that are present in different, unknown proportions. The number of cell types that can be identified is limited by the setting, for instance a pairwise comparison of two tumors can be corrected for a single nuisance. We use sequencing data to estimate both the underlying DNA methylome of each cell type and the composition of each sample. The estimates allow testing cancer cell specific differences and obtaining purity-corrected estimates of the methylomes of each cell type. An overview is shown in Figure 1.

### 2.1 Modeling latent methylation patterns

We assume that there are $n$ (pure) cell types, $j$th of which has an unknown $m$-site methylation pattern $\mathbf{Z}_{:,j} \in \{0,1\}^m$. In a typical case comparing two tumor samples, three hypothesized cell types exist: a single normal cell type, common in both samples, and two cancer cell types, (potentially) unique in each tumor sample (cf. Figure 1). The $m$ sites at which the methylation is modeled can represent either adjacent or distant genomic sites, making the method applicable to whole-genome bisulfite sequencing

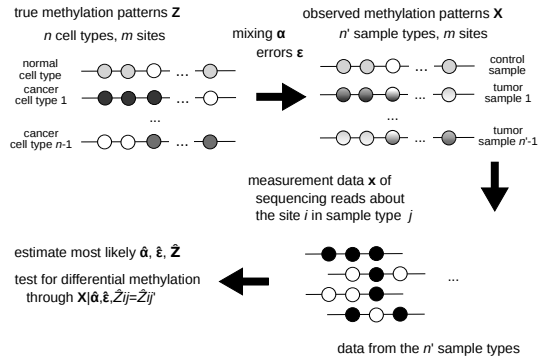


**Fig. 1.** Overview of the proposed methodology. The objective is to identify features of the true methylation patterns of $n$ cell types at $m$ sites. We assume a generative model where each of the $n'$ sample types is a mixture of the $n$ cell types, with unknown proportions and unmodeled additional variations (errors). Sequencing reads from each of the $n'$ sample types are pooled and grouped into $w$-site runs (of CpGs, typically) for local intersite modeling. These data are used to estimate the most likely model, which can be used to test if a pair of cell types feature differential methylation.

(WGBS), targeted bisulfite sequencing (TBS), reduced representation bisulfite sequencing (RRBS) (Harris *et al.*, 2010; Lee *et al.*, 2013), or other sequencing platforms. For methylation, the latent variables $Z_{i,j}$ are binary, but could be, for example, quaternary for detecting DNA mutations.

### 2.2 Modeling observed methylation patterns

A typical tumor sample consists of few major cell types, such as stromal and immune cells in addition to the actual cancer cells of interest (Carter *et al.*, 2012; Aran *et al.*, 2015). Due to this, the methylation patterns that are measured in the impure samples and those of the underlying pure cell types (which are not directly measured) tend to differ.

Each sample is assumed to be a mixture of the $n$ cell types and be further corrupted by random noise (see Figure 1). We use $X_{i,j} \in \{0,1\}$ to denote the random variable corresponding to a methylation read at site $i$ of the $j$th sample type, $\alpha_{k,j}$s to denote the convex mixing parameters and $\epsilon_{k,j}$ to denote the per-site error rate (cf. Figure 1). The mixing parameter $\alpha_{k,j} \in [0,1]$ determines the fraction of cells of type $k$ in the sample $j$, which will be determined by the unknown sample composition. The error rates depend on various factors such as unmodeled variations for the cell type $k$ and sequencing and postprocessing errors, which makes their direct measurement cumbersome, but poses no problem we can estimate them. Here, the errors are assumed to be independent random bit-flips, $\epsilon_{k,j} \in [0,1]$ representing the flip probability, but a more complex error model is possible (see Supplementary material).

In the simple case of comparing two tumor samples, we use a single common error parameter $\epsilon_{k,j} = \epsilon$, three cell types ($n = 3$) as described in the previous section, and three sample types ($n' = 3$): a normal cell sample (control) and two kinds of tumor samples. The mixing is determined by the two mixing parameters $\alpha_{1,1} = \alpha_1$ and $\alpha_{2,2} = \alpha_2$, the tumor purities of each two tumor samples, and the other mixing parameters are implicit: $\alpha_{0,0} = 1$, $\alpha_{k,0} = 0$ (control is pure), $\alpha_{0,j} = 1 - \alpha_{j,j}$ (the impurities in the tumor samples are normal cells) and $\alpha_{2-j+1,j} = 0$ (no crosstalk between the two cancer cell types). In this case, the model density is:

$$\mathbb{P}[\,\mathbf{X}_{I,j} = \mathbf{x}, \mathbf{Z}_{I,:} = \mathbf{z} \,|\, \boldsymbol{\alpha}, \epsilon\,] = \left( \sum_{k=1}^{n} \alpha_{k,j}\, \epsilon^{|\mathbf{x} \neq \mathbf{z}_{:,k}|}\, (1-\epsilon)^{|\mathbf{x} = \mathbf{z}_{:,k}|} \right) \mathbb{P}[\,\mathbf{Z}_{I,:} = \mathbf{z}\,]\,,$$

where $I$ are the site indices covered by the read with values $\mathbf{x}$ in the sample type $j$ and $\mathbf{z}$ are the latent methylation patterns.

Any number of cell and sample types are supported, provided that the problem is identifiable (e.g. it is not possible to decompose a single sample into multiple cell types without further constraints). For example, an arbitrary number of tumors can be compared simultaneously, provided that the normal cells feature a similar methylation pattern in each tumor, or multiple normal cells can be present, provided that an appropriate number of controls are supplied (see Supplementary material). The identifiability problems arising from low sample size are locally mitigated by the co-methylation modeling (see next section).

### 2.3 Co-methylation modeling

The fact that alterations in methylation patterns (Eckhardt *et al.*, 2006) tend to span larger regions, introduces correlations in the adjacent sites $X_{i,j}$ and $X_{i',j}$, which result in incorrect statistical significance and wrong calls. Also, even for RRBS, a single sequencing read will contain multiple adjacent sites ($\sim 3$ to $4$ CpG sites per $100$ basepair read in CpG islands (Illingworth and Bird, 2009)), which introduces correlations in the mixing. While these might not affect the methylation estimates on average, they affect their variance, and hence the significance of differences. While the first type of correlations are amplified in RRBS compared to WGBS, the latter type of correlations are amplified in WGBS.

To capture the correlations, we model $w$-wise joint distributions of the methylation patterns. We found this approach to be the most appropriate, as sufficiently distant sites are expected to be uncorrelated, while the correlation between adjacent sites varies e.g. depending on distance (for RRBS and targeted sequencing) and boundaries of genetic elements (Eckhardt *et al.*, 2006; Lister *et al.*, 2009). Using a larger window size $w$ increases computational effort but has no other disadvantages: if the data lacks dependence, the results equal to those using a smaller window size, so the largest $w$ permitted by computational resources should always be used. Our tests suggests that $w = 2$ is practical for genome-scale analysis, while $w$ of up to about $5$ can be used for smaller datasets, $w$ typically representing the number of adjacent CpG sites in the window.

### 2.4 Model estimation and differential methylation testing

Given the above model, we use sequencing data to simultaneously estimate the distribution $\mathbb{P}[\mathbf{Z}]$ of underlying methylation pattern of each (pure) cell type, the composition $\boldsymbol{\alpha}$ of each tumor sample, and the error rate $\epsilon$.

The estimation is done in maximum likelihood sense (ML; i.e. parameters that most likely generate the data) implemented through a numerical expectation maximization algorithm (see Supplementary material). Direct optimization is infeasible even for moderate number of sites as the complexity is $\mathcal{O}(|\Sigma|^{m\,n})$ where $m$ and $n$ are the number of sites and cell types, respectively, and $|\Sigma| = 2$ is the alphabet size. Meanwhile, our method is exponential time and space in $w\,n$, where $w$ is the window size, but takes linear time (per iteration) and constant memory in $m$, which enables genome-scale analysis. Prior information can be included by modifying the ML objective, such as in cases where the tumor purity estimates are available from other sources.

The estimates allow testing if specifically the cancer cells in the different sample types feature a differential methylation: we derive a p-value through a likelihood ratio test (see Supplementary material). Alternatively, the estimated cell type methylation pattern distributions can be used to obtain most likely methylation patterns for each cell type and "purified" methylation counts (see Supplementary material).

## 3 Results and discussion

### 3.1 Monte Carlo simulations

The use of sequencing data from cancer patients allow estimation of sensitivity to some degree but not specificity. Therefore, a simulation where ground truth is known is important. Here, we employed Monte Carlo simulations based on publicly available WGBS data and compared differential DNA methylation calling under varying tumor purity.

#### 3.1.1 Simulation settings

We obtained sequences of length $m$, for three cell types (one normal and two cancer cells; the simplest setting where multiple tumor samples are compared) by sampling publicly available whole-genome bisulfite sequencing (WGBS) data of immortalized cell lines from The ENCODE Project Consortium (2012) (see Supplementary material). Afterwards, bisulfite reads were generated by sampling the reads of the corresponding WGBS dataset, extracting the methylation signal, and adding random errors. The control sample was generated from the GM12878 cell line, while two tumor samples were generated by mixing the K562 and HepG2 cancer cell lines with the GM12878 line.

Methylation calls were made at a significance level $0.05$ after adjusting for false discovery rate (FDR) using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), following the procedure of Lister *et al.* (2009). An FDR adjustment and the selected significance level strongly influence the number of true versus false calls, but we did not find it to affect our conclusions (not shown).

#### 3.1.2 Performance under varying tumor purity

First, we studied how varying sample purities affects the performance of detecting differentially methylated sites in two tumor samples. We compared our approach with several existing methods: Fisher's exact test (Lister *et al.*, 2009; Pan *et al.*, 2015; Assenov *et al.*, 2016), MOABS (Sun *et al.*, 2014), a mixture-adjusted Fisher's exact test (see Supplementary material), InfiniumPurify (Zheng *et al.*, 2017), and DSS (Feng *et al.*, 2014). In the comparisons, we used our method with and without a control sample and using various window sizes $w$ (in units of CpGs) for the co-methylation modeling. Unlike our method, none of the alternatives can perform control-free differential methylation calling on two tumor samples without known purities, which is a severe limitation, and in the comparisons, they are used with additional information. Most of the alternative methods for detecting differential methylation are unsuitable for a setting comparing two or more heterogeneous tumor samples, as they assume that the samples are pure or feature similar purity (Hebestreit *et al.*, 2013; Hansen *et al.*, 2012; Sun and Yu, 2016; Wang *et al.*, 2016), or cannot compare multiple tumor samples but only tumor versus normal (Zheng *et al.*, 2014).

Figure S1 exemplifies a typical setting with $30\times$ average coverage with $m = 1{,}000$ CpG sites (chromosome 6, 22,333,542–22,469,062 in GRCh38) and sample purities of $\alpha_1 = 0.25$ and $\alpha_2 = 0.75$ and an error rate of $0.05$. As expected, Fisher's exact test performs poorly — not much better than a random guess — due to the different sample purities, while the mixture-adjusted variant can reach an accuracy of about $90\%$. MOABS performs comparably to the Fisher's exact test, and InfiniumPurify and DSS comparably to the mixture-adjusted Fisher's exact test, which is expected as the latter set of methods properly model the tumor composition while MOABS does not. Meanwhile, our method provides higher accuracy than the alternatives over a wide range of thresholds, and generally ranks higher in either (or both) specificity or sensitivity. Using a matched control and higher order co-methylation modeling results in further improvements in the performance, which are unavailable to the alternatives.

We verified that the methods control for false positives as advertised. For this, we performed $500$ simulations with the above settings, but by generating both samples by mixing the GM12878 and K562 cell lines. The results are summarized in Figure S2 and suggest that only Fisher's exact test and MOABS are susceptible of generating excess false positives. Fisher, mix-F, and our methods use a discrete model, causing large p-values to be inaccurate (a staircase-like curve), and the distortions in our method

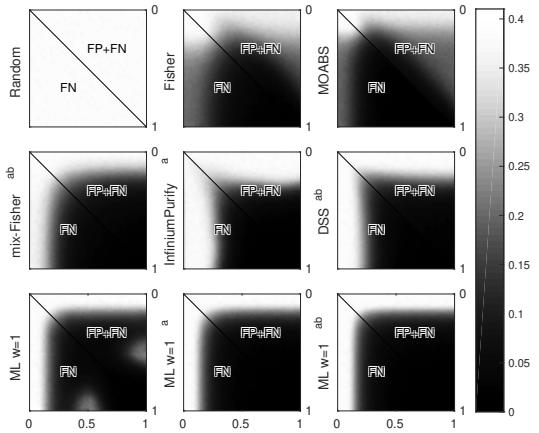


**Fig. 2.** False negatives/false calls with simulated data with varying sample purities. The axes represent tumor purity of the two samples to be compared and the heatmap intensity represents the average false negative rate (FN; lower triangle) and total false calls rate (FN+FP; upper triangle) in 500 simulations. Different panels represent different methods: Fisher's exact test (Fisher), MOABS, mixture-adjusted Fisher's exact test (mix-Fisher), InfiniumPurify, DSS, our maximum likelihood method (ML) with a window size $w = 1$ CpGs. Methods with [a] use a matched control sample, and methods with [b] the true purities. Color version is available in Supplementary material.

likely stem from the asymptotic Wilks' theorem. Regardless, for small significance levels ($< 0.293$), false positives are appropriately controlled for, even though quite conservatively, $\sim 0.02$ for a level of $0.05$.

For a more thorough analysis, we simulated data with varying sample purities in each sample. The results were obtained with $30\times$ average coverage, and the average false call rates of $500$ simulations with $m = 1{,}000$ CpG sites each as a function of the sample purities are visualized in Figure 2. The lower triangle of each heatmap shows the fraction of false negatives, and the upper triangle the total fraction of false calls (false negatives plus false positives; the fraction of false positives ought to be small as they are being controlled for). All methods perform poorly (equal to random) at purities $a_1 = 0$ or $a_2 = 0$. This is expected, as the samples do not contain any information about the methylation patterns of the cancer cells in one (or both) of the samples. Compared to the alternatives, our methods develop these false calls at a later stage. Meanwhile, when the purities are dissimilar, Fisher's exact test and MOABS generate a large amount of false positives, which is the major reason for their unsuitability for these data. Our method, when used without a control, is susceptible to generating false negatives when one of the samples is pure and the other a symmetric mixture (i.e. $a_1 \sim 1$, $a_2 \sim 0.5$ or $a_1 \sim 0.5$, $a_2 \sim 1$). This is due to the fact that in a symmetric, impure sample it is not possible to identify which patterns are from normal and which are from cancer cells, and the independent pure sample cannot aid in the process. This results in significant evidence for the null hypothesis, resulting in very infrequent calls. When a control is provided, there is no such ambiguity, which is also why this issue does not appear with the other methods. Results with higher coverage qualitatively similar, but the accurate range of purities (e.g. $95\%$ accuracy) is increased for each method.

In Figure S3, we show how the performance varies across the simulations. This figure shows the distribution of false calls in the $500$ simulations for each method along the curve $\alpha_1 = 1-\alpha_2$. The results indicate that our methods perform competitively for various degrees of tumor purity, and that providing the control or prior information about the purities offers further advantages. In general, providing a prior results in more consistent performance (lower variance), while providing the control mainly enhances the average performance.

Finally, we verified that the purities are accurately estimated. Figure S4 shows the mean and standard deviation of estimated parameters in the simulations for each method. The results suggests that each method accurately estimates the purity provided that it is above the measurement noise (represented by the error $\epsilon$ in our simulations). InfiniumPurify and our ML method without a control can give highly inaccurate estimates for low purity values, while all methods exhibit a bias in this region. The correlation, shown in the legend of Figure S4, is lowest with InfiniumPurify and highest with our ML method with a control.

#### 3.1.3 Advantages of co-methylation modeling

Next, we show that our method offers even better performance, provided that the sequencing reads are of sufficient length. The alternatives and our method with a window size of $w = 1$ CpGs lack correlation modeling, rendering the information unexploited. This feature allows our method to have good performance even in low coverage settings provided that the reads span over multiple CpG sites.

An example with $30\times$ average coverage in a $1{,}000$ site experiment, with $\alpha = 0.25$, $\alpha_2 = 0.75$, and an error rate of $0.05$, collected from $500$ simulations is summarized in Figure S5. For a read length of exactly $1$ CpG site, the results with $w > 1$ are equal to that with $w = 1$. However, when a single read covers multiple CpG sites, the co-methylation modeling allows for a greater accuracy. Further increases in the window size offer additional improvements in the accuracy. The data that were used as a basis of the simulation features about $4$ CpG sites per read on average, so we are unable to show the advantages beyond this read length, but fully synthetic simulations (not shown) suggest that further improvements are possible for larger read lengths as well.

### 3.2 Ovarian cancer dataset

Next, we applied the methods to detect differential methylation between samples surgically removed from patients diagnosed with high-grade serous ovarian cancer (HGSOC). HGSOC is responsible for more than $40{,}000$ deaths annually in Europe alone and more than $50\%$ of the patients die within five years of diagnosis (Berns and Bowtell, 2012).

#### 3.2.1 Sample description

We used a total of five ovarian cancer tumor samples from three patients, in three comparison settings. Two treatment naive tumors (from initial laparoscopy prior to chemotherapy) were obtained from the peritoneum of patient EOC60 (EOC60-per) and EOC1133 (EOC1133-per), whereas the other three are from interval debulking surgery after three cycles of chemotherapy: one from bowel mesentery of patient EOC60 (EOC60i-meso), and the others from the omentum and ovary of patient EOC868 (EOC868i-ome and EOC868i-ov, respectively). We compared the methylation between EOC60i-meso and EOC60-per (same patient, treatment naive versus interval, different anatomical site), EOC60-per and EOC1133-per (different patient, treatment naive, same anatomical site), and EOC868i-ome and EOC868i-ov (same patient, interval, different anatomical site). A blood sample from patient EOC868 (EOC868-WBC) was used as a normal control where applicable.

The DNA methylomes were profiled using Agilent SureSelect$^{XT}$ Human Methyl-Seq kit (Agilent Technologies, CA, USA) covering $3.7$ M CpGs in $84$ Mb target followed by paired-end sequencing with Illumina HiSeq $2500$ (Illumina Inc., CA, USA). After sequencing, the methylation patterns at the spanned CpG sites were used for further analysis. There were about $8.64$ M such sites, and about $3.07$ M of the sites where both compared samples featured coverage of $5\times$ or more. The $5\times$ coverage filtering was performed to prune out low-quality regions, as controlling the number of false positives is sensitive to these. The average per-site read

coverage before (after) the filtering was about $11.4\times$ ($30.4\times$). Further details are given in Supplementary material.

### 3.2.2 Comparing differential methylation calls

The differential methylation calls within each sample pair were obtained as follows. The acquired p-values for sitewise comparisons were adjusted for false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and calls were made at significance level 0.05, following Lister *et al.* (2009). We used Fisher's exact test (Fisher) and MOABS (Sun *et al.*, 2014), which cannot use a control sample; mixture-adjusted Fisher's exact test (mix-F; see Supplementary material), InfiniumPurify (Zheng *et al.*, 2017), and DSS (Feng *et al.*, 2014), which require a control sample; and our method (ML) with a window size of $w = 1$ or $w = 2$ CpGs, which operates either with or without a control sample. Further, the mixture-adjusted Fisher's test and DSS require external tumor purity estimates, which we derived from the corresponding whole-genome sequencing data (WGS) using ASCAT (Van Loo *et al.*, 2010).

The heatmaps in Figure 3 display the number of mismatches of either type between any two methods. In more than $91\%$ of the calls, any two methods agree. The results indicate that in all cases, the Fisher's exact test based methods share most common calls between themselves rather than with the other methods and vice versa, suggesting that the presence or absence of control is not as important as the choice of the method. Typically, a variant lacking a control makes fewer calls than a variant with one, which is expected as the former lacks statistical power as shown in our simulations. A problem characteristic to our samples is that Fisher, MOABS, mix-F lack confidence to make calls between the samples extracted from a single patient, highlighting the importance of tumor purity adjustment. For mix-F and DSS, an appropriate control enables to avoid this problem, which is available in the EOC868i-ome versus EOC868i-ov comparison. Meanwhile, as InfiniumPurify was designed for microarray data, it does not model the heteroscedasticity resulting from read depth differences, so the calls likely correlate with large effect sizes rather than with overall statistical evidence and thus occur at different sites, which would explain why it differs from all the other methods in most cases. In all comparisons, the ML methods produce novel putative differentially methylated sites regardless whether a control is provided or not. In some instances, putative false positives called by especially the methods lacking tumor purity adjustment are pruned. Our hypothesis is that the tumor purity is so low that the accuracy of the Fisher-based methods starts to deteriorate significantly (cf. Figure 2). Varying the window size results in very few changes. Finally, in the EOC868i-ome versus EOC868i-ov comparison, the ML based methods result in similar results, suggesting that in the case lacking a control, the methylome of the non-cancer cells is accurately estimated. As the other comparisons do not exhibit such property, the blood sample of patient EOC868 is likely an inaccurate representative of the normal cell methylome for these comparisons and the control-free comparison should be preferred. Findings in known ovarian cancer related genes are summarized in Supplementary material.

The computational resources used for the analysis are summarized in Table S1, which indicate that our ML method has a competitive runtime and memory usage when compared to the alternatives. We note that our method has runtime and memory usage that scales linearly with number of CpG sites (see Models and methods), suggesting that even much larger analyses can be done using modest resources.

### 3.2.3 Comparing purity estimates

To validate the accuracy of our purity estimates, we compared our method with ASCAT (Van Loo *et al.*, 2010) and ABSOLUTE (Carter *et al.*, 2012), which estimate the tumor purities from whole-genome sequencing (WGS) data. The purities estimated using ASCAT and ABSOLUTE from

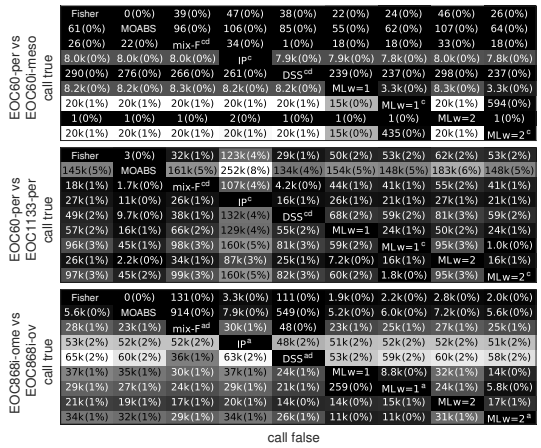


**Fig. 3.** Number/fraction of differential methylation calls in the ovarian cancer samples where the methods differ. The numbers denote the number (fraction) of sites where the method on the row calls for differential methylation and the method on the column does not. As such, the upper triangle shows number of putative false positives assuming the method on the row is correct, while the lower triangle shows putative true negatives assuming the method on the column is correct. Methods: Fisher's exact test (Fisher), MOABS, mixture-adjusted Fisher's exact test (mix-F), InfiniumPurify (IP), DSS, and our ML method with window size of $w = 1$ or $w = 2$ CpGs. Methods with $^{a}$ use a matched control sample, $^{c}$ use an unmatched control, and $^{d}$ use ASCAT purity estimates from WGS data.

the WGS data and InfiniumPurify from the TBS data of our HGSOC samples are shown in Table S2 with the mean and standard deviation in each comparison. The numbers suggest that the samples vary both in tumor purity and in the degree that the purities differ between the two samples, which explains the varying degree of agreement in the differential methylation calls between our ML method and the alternatives.

Next, we compared the above tumor purity estimates to those obtained using our method. The estimates obtained from the methylation data using our method are shown in Table S2, and they tend to follow the other estimates. To test the reliability of our estimates and if they agree with those reported by the other methods, we estimated the parameters from random substrings of the data. For this, $1{,}000$ uniform random $1{,}000$-site regions of consecutive CpG sites were selected for independent parameter estimation. Figure 4 shows the distribution of estimated parameters for the random substrings of each tumor sample using the methylation data and our method. To test the agreement between the estimates produced by ASCAT, ABSOLUTE, or InfiniumPurify and the estimate by our method, we computed the p-value of obtaining each estimate under the null hypothesis that they follow the distribution of distances from our estimate specified by the random substring estimates, as shown in Table S2. To conclude, we found no evidence that the estimates produced by our method are in disagreement with any of the ASCAT or InfiniumPurify estimates. However, we found some evidence that the ABSOLUTE estimates for the EOC868i-ome and EOC868i-ov samples might differ from those produced by the other methods (p-values $\sim 0.03$ when compared with the ML method), which might suggest ABSOLUTE performs better than ASCAT, InfiniumPurify, and our method in low ($< 20\%$) tumor purity settings. In addition, the results indicate that the estimated error parameter is small (less than $0.10$ in $77\%$ of the cases) in each case (with no constraints), suggesting that the mixture model explains majority of the data well.

### 3.2.4 Differential methylation as a predictor of differential expression

To test whether the differential methylation calls produced by our method are more accurate than those of the reference methods, we tested whether our method better predicts gene expression data. For more accurate DNA methylome quantification, we expect to see a stronger correlation

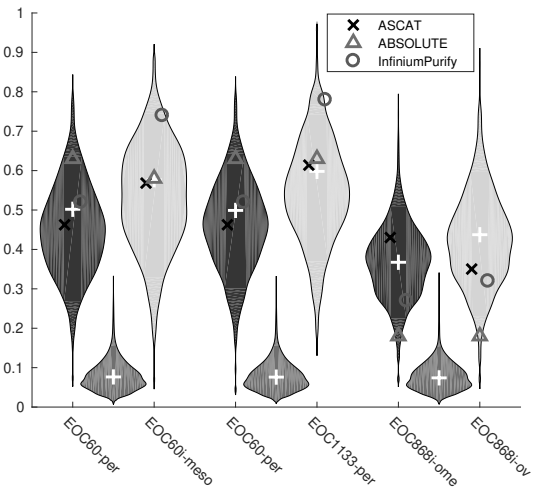


**Fig. 4.** Parameter estimates and an estimate of their variability. Parameter estimates from $1{,}000$ pieces of $1{,}000$ neighboring sites (violins), the estimates from the whole genome (white crosses), and the tumor purity estimates from WGS data using ASCAT, ABSOLUTE, and from methylation data using InfiniumPurify. Darker and lighter violins correspond to the purities of the compared samples, and the intermediate the (common) error rate.

between the predicted differential methylation in promoter regions and the corresponding difference in gene expression.

For this, we quantified the Spearman's rank correlation between the RNA expression ratio and the difference in the DNA methylation in the compared samples. For each method, the difference in the average methylation corrected using the ASCAT purity estimate and the average methylation of the control were used (i.e. solving $y = (1-\alpha_j)\, y_0 + \alpha_j\, m$ for $m$ where $y_0, y$ are the observed average methylation of the control and the tumor sample, respectively, $\alpha$ is the purity, and $m$ is the "purified" average methylation), setting the prediction to zero where no call was made. Similarly, the expression values were corrected using the ASCAT purity estimate and a control pooled from all the samples (median). Rank correlation ensures that batch effects and nonlinear relationships between the covariates play no role. The expression data were obtained by analyzing the corresponding RNA-seq data sequenced from each tumor sample.

The correlation was quantified in the $[-1500, +500)$ region (i.e. the promoter and the first exon area; see Figure S6) about the transcription start sites of known genes (extracted Jan 2017 from Ensembl (Yates *et al.*, 2016)). Hypermethylation in these areas often results in suppressed expression (Witte *et al.*, 2014). Outside of the region, no significant genome-wide correlation was found. Coincidentally, the effects of global hypomethylation and more focal alterations are expected to be less visible in a genome-wide correlation.

To compare two methods, we quantified the correlation for CpG sites called by either of the two methods. A t-test was used to determine the presence of significant correlation, and Fisher's transform and a z-test to determine if there is a significant difference between two correlation values (i.e. if two potentially non-zero correlation values differ significantly). The results comparing Fisher's exact test, MOABS, InfiniumPurify, and DSS to our ML method with a control are shown in Table 1. In each comparison, our ML predictor results in a significant anticorrelation ($\sim -8$ to $-12\%$) between differential promoter methylation and the expression ratio in each sample pair. The results for DSS suggest a similar pattern. For others, significant correlation only exist in some samples. More importantly, our ML predictor results in a stronger correlation, which was found to be significant except against DSS (at significance level $0.05$), suggesting that our method results in a more accurate recovery of the methylome.

Table 1. Rank correlation between the predicted differential methylation and the differential expression ratio

| Sample A/B | Method | Corr - | $P_{-,0}$ | Corr ML | $P_{ML,0}$ | $P_{-,ML}$ |
|---|---|---|---|---|---|---|
| 60-per /60i-meso | F | $-3.6\%$ | $0.04$ | $-9.5\%$ | $<10^{-7}$ | $0.02$ |
| 60-per /1133-per | F | $-6.7\%$ | $<10^{-26}$ | $-10.6\%$ | $<10^{-65}$ | $<10^{-5}$ |
| 868i-ome/868i-ov | F | $-2.4\%$ | $0.08$ | $-13.1\%$ | $<10^{-20}$ | $<10^{-7}$ |
| 60-per /60i-meso | MOABS | $-2.6\%$ | $0.13$ | $-9.5\%$ | $<10^{-7}$ | $<10^{-2}$ |
| 60-per /1133-per | MOABS | $-6.9\%$ | $<10^{-38}$ | $-9.0\%$ | $<10^{-65}$ | $<10^{-2}$ |
| 868i-ome/868i-ov | MOABS | $-1.7\%$ | $0.20$ | $-12.6\%$ | $<10^{-20}$ | $<10^{-8}$ |
| 60-per /60i-meso | IP | $-3.1\%$ | $0.04$ | $-8.5\%$ | $<10^{-7}$ | $0.01$ |
| 60-per /1133-per | IP | $-5.8\%$ | $<10^{-16}$ | $-11.6\%$ | $<10^{-65}$ | $<10^{-9}$ |
| 868i-ome/868i-ov | IP | $-3.9\%$ | $<10^{-04}$ | $-9.4\%$ | $<10^{-20}$ | $<10^{-4}$ |
| 60-per /60i-meso | DSS | $-4.7\%$ | $<10^{-2}$ | $-9.5\%$ | $<10^{-7}$ | $0.05$ |
| 60-per /1133-per | DSS | $-10.5\%$ | $<10^{-64}$ | $-10.5\%$ | $<10^{-64}$ | $0.99$ |
| 868i-ome/868i-ov | DSS | $-10.1\%$ | $<10^{-30}$ | $-8.2\%$ | $<10^{-20}$ | $0.13$ |
| 60-per /60i-meso | ML $w=1$ | $-8.4\%$ | $<10^{-05}$ | $-9.4\%$ | $<10^{-7}$ | $0.67$ |
| 60-per /1133-per | ML $w=1$ | $-12.1\%$ | $<10^{-65}$ | $-12.1\%$ | $<10^{-65}$ | $0.99$ |
| 868i-ome/868i-ov | ML $w=1$ | $-9.6\%$ | $<10^{-12}$ | $-12.4\%$ | $<10^{-20}$ | $0.13$ |

The blocks contain comparison between the reference method and our maximum-likelihood (ML) method with window size of $w = 2$. The correlations were computed from the sites where one (or both) of the compared methods make a call. The table lists the correlation for the reference method (Corr -) and the ML method (Corr ML), p-values for the hypotheses that these correlations are zero ($P_{-,0}$, $P_{ML,0}$), and p-values for the hypotheses that these correlations are equal ($P_{-,ML}$).

### 3.3 Lymphoma dataset

We also applied the methods on samples collected from diffuse large B-cell lymphoma (DLBCL) patients (see Supplementary material). Again, the results between the methods largely agree (for $>83\%$ of the calls, any two methods agree; Figure S9, Figure S11–Figure S13). However, provided that our methods are more accurate — as suggested by our simulations and the ovarian cancer analysis — our methods allow ruling out a sizable fraction of false positive calls, and, when used with a control sample, suggest novel findings which the previous methods are unable to identify. Interestingly, the nature of the control sample does not seem to drastically affect identifying the false positives, so even an unmatched control derived from a different tissue, other patients, or from a database is suitable to gain some improvements. The purity estimates are consistent with those of ASCAT, ABSOLUTE, and InfiniumPurify (Figure S10).

## 4 Conclusion

We have introduced a method that estimates differential methylation between multiple cancer samples featuring varying, unknown tumor purity. The method does not require a prior estimate of the sample purities, nor the methylome of a normal sample, but estimates these in the process. If purity estimates or normal samples are available, they can be used to the improve the estimator accuracy.

The developed method is expected to be of paramount value toward personalized medicine applications, where prognosis, treatment decisions, and response follow-up need to be done using multiple samples, harvested from multiple locations and time points at the individual patient level. Such approach is fundamental for studying heterogeneous diseases, where the comparison can only be done at the patient or small, stratified subgroup level. Our method is the first that allows comparison between tumor and healthy tissue or blood samples, between primary and metastatic tumor samples, or between samples of different patients under the condition where the sample purities vary and a reliable control is absent.

We used Monte Carlo simulations to demonstrate that the method can operate for a wide range of purities, unlike Fisher's exact test, MOABS, or other methods which do not model the sample composition. Fisher's exact test and few other methods like DSS can be adjusted for tumor purities but, unlike our method, the adjustment requires prior knowledge of the purities and a reliable control sample. Our method can also exploit such information, but also of the co-methylation of closely located sites, which is a recognized phenomenon (Lister *et al.*, 2009; Eckhardt *et al.*, 2006) but is neglected by a sitewise Fisher's exact test and most alternatives. In general, regardless of the setting, our method outperforms all the compared alternatives for a wide range of sensitivity versus specificity.

We applied the method on targeted bisulfite sequencing data from ovarian cancer patients. The results largely agree between the methods, but our method suggested up to $5\%$ novel differentially methylated sites, which the previous methods were unable to identify, and allowed ruling out some false positives. The superior accuracy of our method was confirmed by predicting RNA expression data analyzed with independent methods, and the tumor purity estimates were validated using independent methods using both whole-genome sequencing data (ASCAT and ABSOLUTE) and the methylation data (InfiniumPurify). Thus, we expect that the performance of our method for tumor purity estimation is comparable to other methods as well. The methods were also applied on DLBCL patient data, demonstrating that our method adapts well to different cancer types and a wide range of tumor purities, and it works robustly with controls derived from various sources. The analyses also demonstrate the method can be deployed in genome scale (up to $\sim 450$ M samples per comparison) and works robustly both with and without a normal cell control.

As genome-wide profiling of DNA methylation has been enabled only recently, computational methods are being developed in order to analyze these data in a meaningful sense. Our method is a significant contribution to this effort for several reasons: First, as the method can estimate all model parameters simultaneously, it requires no additional measurements for configuration parameters. Second, the method can operate either with or without a control sample, which is critical as a control is often unavailable or unreliable. Third, more accurate methods are necessary to elucidate less prominent differences; or, fewer data are needed for equivalent power. Due to these and the important role of DNA methylation, we expect our method to greatly benefit research on cancer and other complex diseases.

The methodology generalizes for estimating differences in any sequences, such as mutations in unconverted DNA or copy-number variations, and already support comparison of multiple samples and more complex sample compositions than used here. Taken together, we believe that our methods enjoy wide applicability in analyzing and comparing measurement data from various sequencing-based platforms.

## Funding

This work was supported financially by the Academy of Finland (Center of Excellence in Cancer Genetics Research and OVCURE), European Union's Horizon 2020 research and innovation programme under grant agreement No. 667403, and Finnish Cancer Organizations. CSC – IT Center for Science is acknowledged for providing computing resources.

## References

Altman, A. D., Nelson, G. S., Ghatage, P., *et al.* (2013). The diagnostic utility of TP53 and CDKN2A to distinguish ovarian high-grade serous carcinoma from low-grade serous ovarian tumors. *Mod. Pathol.*, **26**(9), 1255–1263.

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.

Assenov, Y., Muller, F., Lutsik, P., *et al.* (2016). Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**(11), 1138–1140.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, **57**(1), 289–300.

Berns, E. M. J. J. and Bowtell, D. D. (2012). The changing view of high-grade serous ovarian cancer. *Cancer Res.*, **72**(11), 2701–2704.

Carter, S. L., Cibulskis, K., Helman, E., *et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**(5), 413–421.

Ciriello, G., Miller, M. L., Aksoy, B. A., *et al.* (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**(10), 1127–1133.

Eckhardt, F., Lewin, J., Cortese, R., *et al.* (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**(12), 1378–1385.

Feng, H., Conneely, K. N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucl. Acids Res.*, **42**(8), e69.

Frommer, M., McDonald, L. E., Millar, D. S., *et al.* (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.*, **89**(5), 1827–1831.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, **144**(5), 646–674.

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**(10), R83.

Harris, R. A., Wang, T., Coarfa, C., *et al.* (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**(10), 1097–1105.

Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**(13), 1647–1653.

Houseman, E. A., Accomando, W. P., Koestler, D. C., *et al.* (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinf.*, **13**(1), 86.

Illingworth, R. S. and Bird, A. P. (2009). CpG islands – 'a rough guide'. *FEBS Lett.*, **583**(11), 1713–1720.

Lee, E.-J., Luo, J., Wilson, J. M., and Shi, H. (2013). Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer Lett.*, **340**(2), 171–178.

Lister, R., Pelizzola, M., Dowen, R. H., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**(7271), 315–322.

Pan, H., Jiang, Y., Boi, M., *et al.* (2015). Epigenomic evolution in diffuse large B-cell lymphomas. *Nat. Commun.*, **6**, 6921.

Shen, H. and Laird, P. W. (2013). Interplay between the cancer genome and epigenome. *Cell*, **153**(1), 38–55.

Sun, D., Xi, Y., Rodriguez, B., *et al.* (2014). MOABS: Model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Sun, S. and Yu, X. (2016). HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Stat. Appl. Genet. Mol. Biol.*, **15**(1), 55–67.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Timp, W. and Feinberg, A. P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer*, **13**(7), 497–510.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., *et al.* (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.*, **107**(39), 16910–16915.

Wang, X., Gu, J., Hilakivi-Clarke, L., *et al.* (2016). Dm-bld: differential methylation detection using a hierarchical bayesian model exploiting local dependency. *Bioinformatics*, **33**(2), 161–168.

Wei, S. H., Balch, C., Paik, H. H., *et al.* (2006). Prognostic DNA methylation biomarkers in ovarian cancer. *Clin. Cancer Res.*, **12**(9), 2788–2794.

Witte, T., Plass, C., and Gerhauser, C. (2014). Pan-cancer patterns of DNA methylation. *Genome Med.*, **6**(8), 66.

Yang, Z., Jones, A., Widschwendter, M., and Teschendorff, A. E. (2015). An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol.*, **16**, 140.

Yates, A., Akanni, W., Amode, M. R., *et al.* (2016). Ensembl 2016. *Nucl. Acids Res.*, **44**(D1), D710–D716.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., *et al.* (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.

Zheng, X., Zhao, Q., Wu, H.-J., *et al.* (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.*, **15**(8), 419.

Zheng, X., Zhang, N., Wu, H.-J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of dna methylation data from cancer studies. *Genome Biol.*, **18**(1), 17.