PLOS ONE

RESEARCH ARTICLE

Statistical methods for classification of 5hmC levels based on the Illumina Inifinium HumanMethylation450 (450k) array data, under the paired bisulfite (BS) and oxidative bisulfite (oxBS) treatment

Alla Slynko[1]*, **Axel Benner**[2]

1 Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada, **2** Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany

* alla.a.slynko@gmail.com

OPEN ACCESS

Citation: Slynko A, Benner A (2019) Statistical methods for classification of 5hmC levels based on the Illumina Inifinium HumanMethylation450 (450k) array data, under the paired bisulfite (BS) and oxidative bisulfite (oxBS) treatment. PLoS ONE 14(6): e0218103. https://doi.org/10.1371/journal.pone.0218103

Editor: Robert Dante, Centre de Recherche en Cancerologie de Lyon, FRANCE

Received: December 4, 2018

Accepted: May 27, 2019

Published: June 13, 2019

Data Availability Statement: https://zenodo.org/record/2639285#.XLYzNKZS_XE.

Funding: This project was supported by grants from the German Federal Ministry of Education and Research (01ER1505a, 01ER1505b) and the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany. The funders did not play any role in the data analysis, decision to publish, or preparation of the manuscript.

Abstract

Hydroxymethylcytosine (5hmC) methylation is a well-known epigenetic mark that is involved in gene regulation and may impact genome stability. To investigate a possible role of 5hmC in cancer development and progression, one must be able to detect and quantify its level first. In this paper, we address the issue of 5hmC detection at a single base resolution, starting with consideration of the well-established 5hmC measure $\Delta\beta$ and, in particular, with an analysis of its properties, both analytically and empirically. Then we propose several alternative hydroxymethylation measures and compare their properties with those of $\Delta\beta$. In the absence of a gold standard, the (pairwise) resemblance of those 5hmC measures to $\Delta\beta$ is characterized by means of a similarity analysis and relative accuracy analysis. All results are illustrated on matched healthy and cancer tissue data sets as derived by means of bisulfite (*BS*) and oxidative bisulfite converting (*oxBS*) procedures.

Introduction

DNA methylation is known to play a crucial role in the development of diseases such as diabetes, schizophrenia, and some forms of cancer; for details see, e.g., [1–7] and references therein. In order to address the possible impact of DNA methylation on the various biological functions and processes, an entire strand of extensive biological, bioinformatical, and statistical analyses has been developed in the past years. Some of those analyses, most relevant for our setting, were discussed in [8–15]. A substantial part of the methods introduced in those analyses aims at quantifying the actual level of DNA methylation, in particular on a single nucleotide resolution in genomic DNA.

At some point, this research indicated that the obtained DNA methylation level, sometimes referred to as "total DNA methylation" [16, 17], can be split, inter alia, into 5-hydroxymethylcytosine (5hmC) and 5-methylcytosine (5mC) components, with 5mC playing an important

Competing interests: The authors have declared that no competing interests exist.

role in gene silencing and genome stability [18]. The second component, 5hmC methylation, was first discovered in 2009 as another form of cytosine modification [19–21]. Since then, its function as an intermediate in active DNA demethylation and an important epigenetic regulator of mammalian development which is strongly associated with genes and regulatory elements in the genome, as well as its role as a possible epigenetic mark impacting genome stability has come into the spotlight [16, 18, 22–38]. At that point, the questions concerning reliable identification and accurate quantification of 5hmC levels emerged.

Until now, a number of techniques for the quantification of 5hmC levels have been established [16–18, 24, 28, 31, 39–41]. Two key techniques to be named here are the TET-assisted bisulfite (*TAB*) technique and the oxidative bisulfite (*oxBS*) technique. The *TAB* technique is based on the conversion of 5mC to 5hmC in mammalian DNA by means of TET emzymes [17, 31]. When using the *oxBS* technique, 5hmC methylation levels can be obtained by means of the paired bisulfite sequencing (*BS*) and oxidative bisulfite sequencing (*oxBS*) procedures [18]. In particular, since the *BS* procedure can only differentiate between methylated and unmethylated cytosine bases, and cannot discriminate between 5mC and 5hmC, the *oxBS* procedure must be applied, in order to determine the level of 5hmC at a considered nucleotide position. This procedure yields Cs only at 5mC sites while oxidating 5hmC to 5-formylcytosine (5fC) and later converting them to uracil. As a result, an amount of 5hmC at each particular nucleotide position can be determined as the difference between the *oxBS* (which identifies 5mC) and the *BS* (which identifies 5mC+5hmC) readouts. In the present paper, all obtained results are illustrated on paired *BS* and *oxBS* data.

In order to quantify the 5hmC level in the context of the *oxBS* technique, and, in particular, to identify a given CpG site as being either hydroxymethylated or non-hydroxymethylated, the following quantity was introduced in [41]

$$\Delta\beta^{oxBS} = \beta_{BS} - \beta_{oxBS} = \frac{M_{BS}}{M_{BS} + U_{BS} + 100} - \frac{M_{oxBS}}{M_{oxBS} + U_{oxBS} + 100}. \quad (1)$$

Here, M denotes the intensity of the methylated allele, U is the intensity of the unmethylated allele, β_{BS} is the methylation level obtained from the *BS* method, and β_{oxBS} is the methylation level derived by means of the *oxBS* method. As stated in [31, 41], the quantity $\Delta\beta^{oxBS}$ computed for each single CpG and sample can be interpreted as a "measure of hydroxymethylation" and "a reflection of the 5hmC level at each particular probe location". This measure can then be applied in the screening step so as to exclude from further analysis those CpGs that do not appear to be hydroxymethylated.

In [42], the authors introduced a related quantity Δm^{oxBS}, defined as a difference of the corresponding m-values [13], to be another measure for identification and quantification of the 5hmC levels. However, our discussion in S1 Appendix shows that in the context of the 5hmC identification both measures, $\Delta\beta^{oxBS}$ and Δm^{oxBS}, flag exactly the same cytosines as being substantially hydroxymethylated and thus can be used *interchangeably*.

Due to its definition, $\Delta\beta^{oxBS}$ in (1) can take values between -1 and 1, with negative values of $\Delta\beta^{oxBS}$ representing "false differences in methylation score between paired *BS*-only and *oxBS* data sets" and being interpreted as a "background noise" [41].

While applying $\Delta\beta^{oxBS}$ for the identification of substantially hydroxymethylated cytosines, the issue of an appropriate $\Delta\beta^{oxBS}$ *threshold* arises; such threshold can be applied "to identify a probe-set of substantially hydroxymethylated cytosines". In [41], the threshold for $\Delta\beta^{oxBS}$ has been set to 0.3 or 30%. However, it is not evident, whether such threshold can be applied for any given data set or should be specified for each particular setting.

This paper is organized as follows. First, we address the applicability of the 5hmC measure $\Delta\beta^{oxBS}$ (in the following notation just $\Delta\beta$) for detection of hydroxymethylated CpGs and then indicate several limitations of this measure by discussing its properties, both analytically and on data sets. Further, we propose several alternative hydroxymethylation measures which can also be applied for the 5hmC identification and compare their properties and resemblance with those of $\Delta\beta$. Relative accuracy and resemblance of all three considered 5hmC measures are discussed numerically, under the assumption that no gold standard is available. All data analyses were performed on 38 matched samples, with cancer and healthy tissue available for each sample.

Discussion

On the applicability of $\Delta\beta$ for 5hmC detection

According to [8], for a given methylated and unmethylated intensities M and U, the methylation level of the particular probe can be described by the *methylation proportion*

$$\beta = \frac{M}{M + U + 100}. \tag{2}$$

Thus, the 5hmC measure $\Delta\beta^{oxBS}$ in (1) is just the difference of two methylation proportions as derived from *BS* and *oxBS* treatment, respectively. This simple definition, while appearing to be plausible at first, nevertheless leads to a number of ambiguities as discussed below.

The first ambiguity arising from (1) concerns the application of $\Delta\beta$ as a measure for the identification of hydroxymethylated CpGs, and, in particular, its adequate interpretation as such. Even if both components in the difference (1) do represent the respective methylation proportions for *BS* and *oxBS* data, these proportions are evidently calculated on two different bases: the proportion β_{BS} represents the methylation proportion based on the global *BS* intensity $M_{BS} + U_{BS}$, whereas the proportion β_{oxBS} represents the methylation proportion based on the global *oxBS* intensity $M_{oxBS} + U_{oxBS}$. Thus, a direct comparison of these two proportions is difficult to justify and, as a result, the interpretation of $\Delta\beta$ as "a reflection of the 5hmC level at each particular probe" suggested in [41] is not well founded.

Further, while identifying hydroxymethylated CpGs in the context of the screening step, the outcomes of $\Delta\beta$ are interpreted as follows [41]: Positive values of $\Delta\beta$ are taken as an indicator for a substantial 5hmC level and "represent potential sites of 5hmC", whereas small values of $\Delta\beta$ should indicate no or only nonsubstantial hydroxymethylation levels. Negative values of $\Delta\beta$ are considered as resulting from background noise; for the 5hmC measure Δm, the same view is shared in [42]. To analyze this interpretation, let us first refer to Fig 1. As the left-hand panel of that figure shows, all ten simulated data points $s_1, s_2, \ldots, s_{10}$ satisfy both conditions

$$M^i_{BS} < M^i_{oxBS} \text{ and } U^i_{BS} < U^i_{oxBS}, \qquad i = 1, 2, \ldots, 10 \tag{3}$$

simultaneously which intuitively should be interpreted as "no substantial 5hmC level observed". Nevertheless, the condition $\Delta\beta > 0$ holds for each of these ten data points as well; see the right-hand panel of Fig 1 for an illustration.

Further, the left-hand panel of Fig 2 introduces another ten simulated data points $s_1, s_2, \ldots s_{10}$ that satisfy both

$$M^i_{BS} > M^i_{oxBS} \text{ and } U^i_{BS} > U^i_{oxBS}, \qquad i = 1, 2, \ldots, 10. \tag{4}$$

At the same time, the condition $\Delta\beta < 0$ holds for each of $s_1, s_2, \ldots s_{10}$ as well; see the right-hand panel of Fig 2 for an illustration. Thus, even though the data points $s_1, s_2, \ldots s_{10}$ in Fig 2

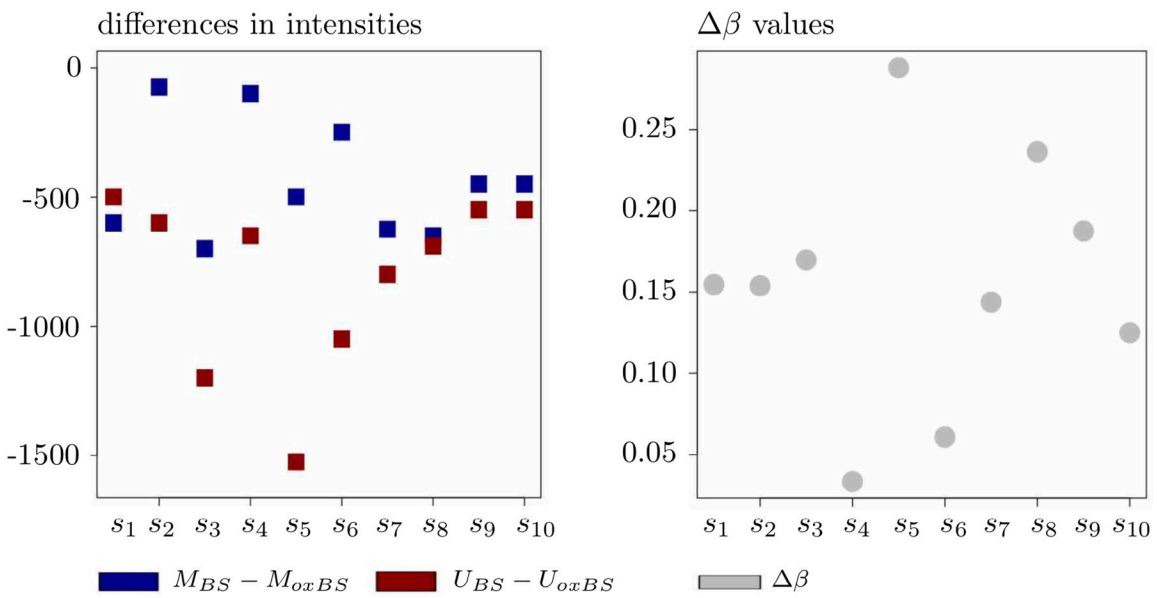

Fig 1. On the interpretation of $\Delta\beta$ as a 5hmC measure in case with $\Delta\beta > 0$. Negativity of the differences on the left-hand panel implies that none of the data points $s_1, s_2, \ldots s_{10}$ shows any substantial 5hmC level, but, due to $\Delta\beta > 0$, all these points will nevertheless be flagged by $\Delta\beta$ as being hydroxymethylated.

https://doi.org/10.1371/journal.pone.0218103.g001

actually appear to exhibit a substantial 5hmC level due to their *BS* intensities exceeding their *oxBS* intensities, they will not be selected by the measure $\Delta\beta$ as being hydroxymethylated.

One of the main advantages of the measure β, which has definitely contributed to its common application as a methylation measure, is its intuitive interpretation as an approximation

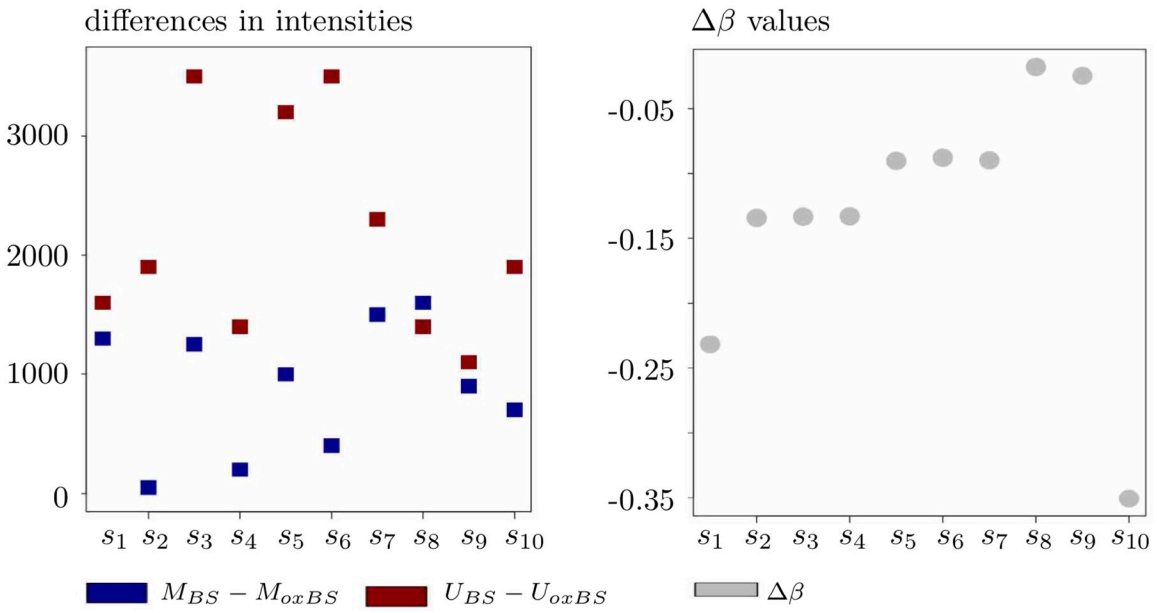

Fig 2. On the interpretation of $\Delta\beta$ as a 5hmC measure in case with $\Delta\beta < 0$. Due to the positivity of the differences on the left-hand panel, all ten data points $s_1, s_2, \ldots, s_{10}$ appear to exhibit a substantial level of 5hmC, whereas the right-hand panel shows negative $\Delta\beta$ values.

https://doi.org/10.1371/journal.pone.0218103.g002

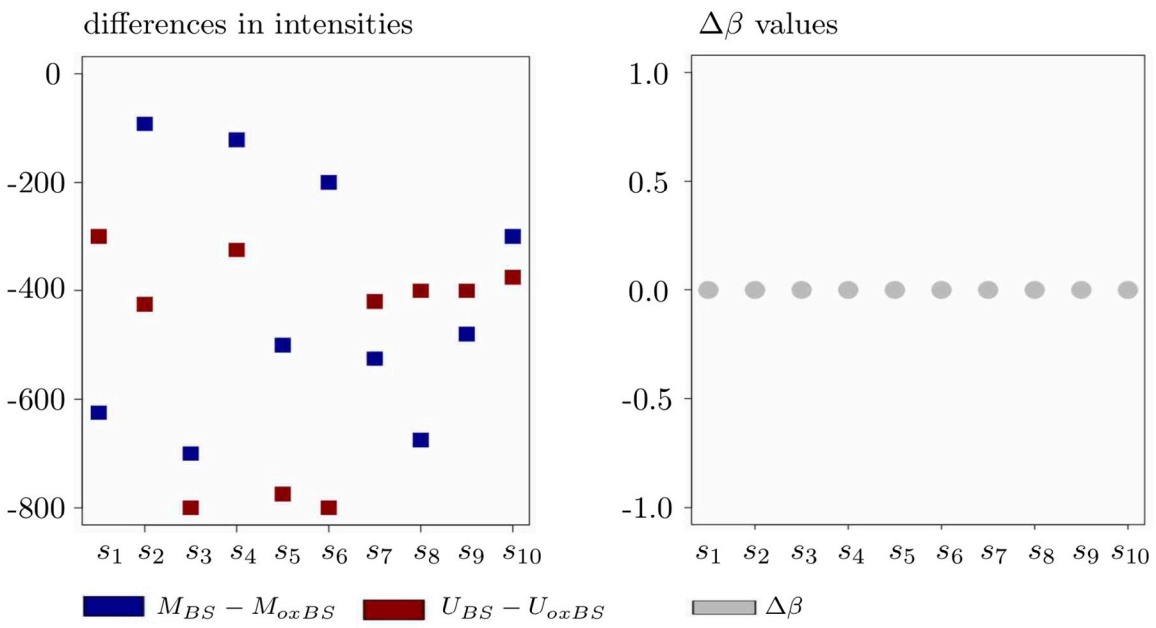

Fig 3. On the interpretation of $\Delta\beta$ as a 5hmC measure in case with $\Delta\beta = 0$. Negativity of the differences on the left-hand panel implies that none of the data points should show any substantial 5hmC level.

https://doi.org/10.1371/journal.pone.0218103.g003

of the *percentage of methylation* [13]; thereby $\beta = 0$ indicates unmethylated probes and $\beta = 1$ denotes fully methylated probes. Unfortunately, this interpretation does not carry over to the measure $\Delta\beta$. Indeed, in (1) the condition $\Delta\beta = 0$ solely implies

$$\frac{M_{BS}}{M_{oxBS}} = \frac{U_{BS} + 100}{U_{oxBS} + 100} \tag{5}$$

and it is unclear how this last equality should be interpreted in terms of the observed 5hmC level. In particular, Fig 3 demonstrates that we can obtain $\Delta\beta = 0$ in cases with "no substantial 5hmC level observed", i.e., in cases where the conditions

$$M^i_{BS} < M^i_{oxBS} \text{ and } U^i_{BS} < U^i_{oxBS}, \qquad i = 1, 2, \ldots, 10 \tag{6}$$

hold. Similar results can be derived in cases with "a substantial 5hmC level observed", i.e., in cases with

$$M^i_{BS} > M^i_{oxBS} \text{ and } U^i_{BS} > U^i_{oxBS}, \qquad i = 1, 2, \ldots, 10. \tag{7}$$

Altogether, our analyses of the conditions $\Delta\beta > 0$, $\Delta\beta = 0$, and $\Delta\beta < 0$ show that their interpretations as indicators for substantial hydroxymethylation, no hydroxymethylation, and background noise may become problematic in certain situations.

Another ambiguity arising from (1) is related to the choice of the number 100 in the denominators $M_{BS} + U_{BS} + 100$ and $M_{oxBS} + U_{oxBS} + 100$ of the expression (1) for $\Delta\beta$. This choice seems to stem from the practical convention in the definition of β values [13], and just being transferred at the definition of $\Delta\beta$ [31, 41]. As a matter of fact, there is no strong reason why the correction term 100 in the denominator of (2) should not be replaced with any other value $\alpha > 0$. In fact, such replacement would lead to the following more general definition of

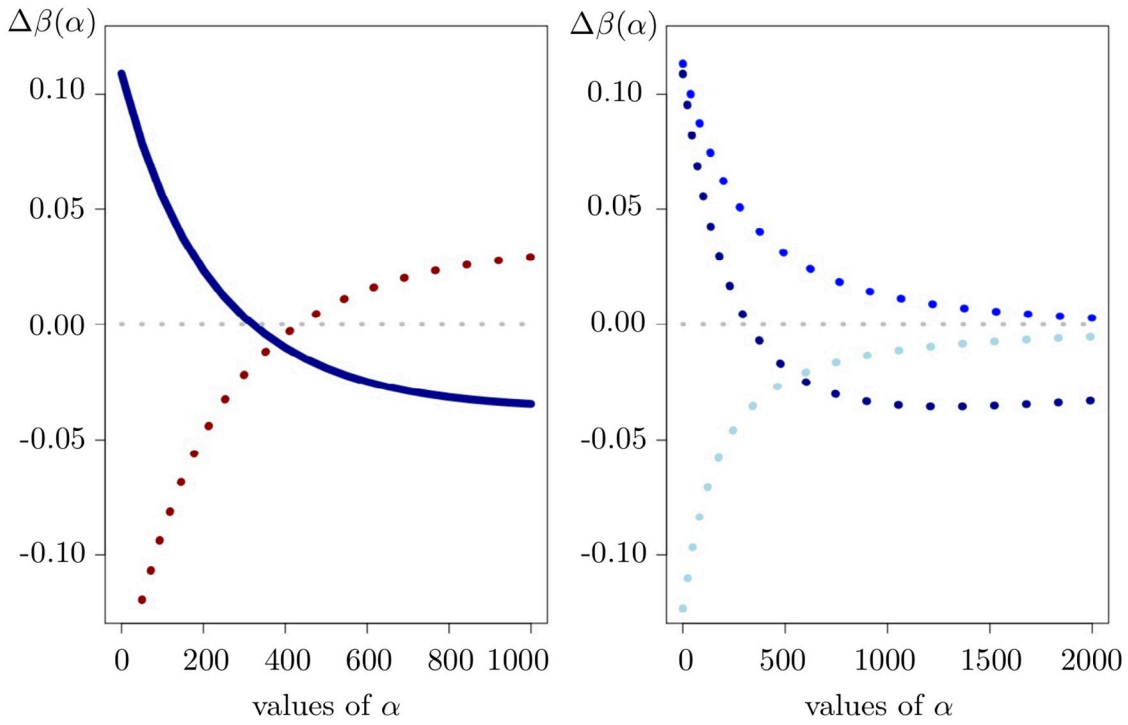

Fig 4. Sign change and convergence of the 5hmC measure $\Delta\beta(\alpha)$. The left-hand panel: $\Delta\beta(\alpha)$ changing its sign from positive to negative (the dark blue curve, healthy tissue) and from negative to positive (the dark red dotted curve, cancer tissue) as α increases. The result refers to a given CpG (*cg*00050873) and sample (*sample 7*). The right-hand panel: Convergence of $\Delta\beta(\alpha)$ for healthy tissue, a given sample (*sample 7*) and three CpGs (*cg*00050873, *cg*05480730, *cg*10698069).

https://doi.org/10.1371/journal.pone.0218103.g004

the methylation proportion

$$\beta(\alpha) = \frac{M}{M + U + \alpha}, \quad \text{with} \quad \alpha > 0. \tag{8}$$

While one can safely argue that the actual choice of the parameter α is not crucial for the interpretation of the methylation proportion $\beta(\alpha)$ itself [13], this choice may become critical when using the sign of the measure

$$\Delta\beta(\alpha) = \beta_{BS}(\alpha) - \beta_{oxBS}(\alpha) = \frac{M_{BS}}{M_{BS} + U_{BS} + \alpha} - \frac{M_{oxBS}}{M_{oxBS} + U_{oxBS} + \alpha}, \tag{9}$$

as an indicator for hydroxymethylation in the screening step. In particular, under certain conditions, the sign of $\Delta\beta(\alpha)$ can change from positive to negative or vice versa as α varies; see the left-hand panel of Fig 4 as well as Fig A in S1 Appendix for an illustration.

Further, Fig 5 shows changes in the density of $\Delta\beta(\alpha)$ as well as the percentage of CpGs (for each given sample) where $\Delta\beta(\alpha)$ may change its sign as α increases.

In view of such dependence of $\Delta\beta(\alpha)$ on the choice of α, a question concerning the possible impact of this choice on the percentage of CpGs satisfying the condition $\Delta\beta(\alpha) > 0$ and thus identified as being hydroxymethylated at the end of the screening step arises.

Alternative 5hmC measures

One of the limitations of the 5hmC measure $\Delta\beta(\alpha)$ we discussed in the previous section concerns its interpretation and robustness with respect to the choice of the correction term α. To

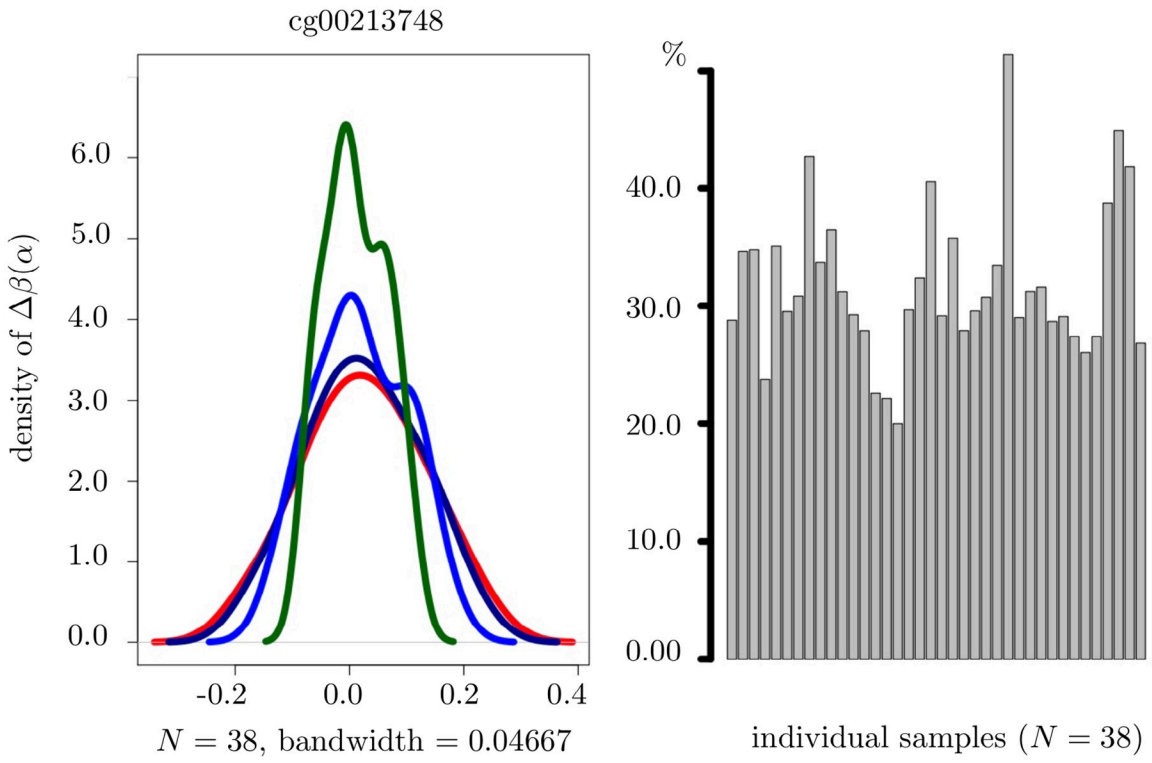

Fig 5. Density of $\Delta\beta(\alpha)$ and the percentage of CpGs, where $\Delta\beta(\alpha)$ may change its sign for varying values of α. The left-hand panel: Density of $\Delta\beta(\alpha)$, for a given CpG (*cg*00213748) and $\alpha = 0$, 100, 500 and 2000 (the red, the dark blue, the blue and the dark green curves, respectively). The right-hand panel: The percentage of CpGs, where $\Delta\beta(\alpha)$ may change its sign for varying values of α. All results were computed on cancer tissue and across all 38 samples.

https://doi.org/10.1371/journal.pone.0218103.g005

overcome this limitation, we now introduce two alternative measures which can be used in the screening procedure while indicating CpGs with a substantial level of 5hmC; the basic properties of these measures are discussed in S2 Appendix.

We start our analysis by considering the behavior of $\Delta\beta(\alpha)$ (and also $\Delta m(\alpha)$ as discussed in S2 Appendix) for increasing values of α. As follows from (9), $\Delta\beta(\alpha)$ vanishes as α increases; see the right-hand panel of Fig 4 for an illustration. This convergence result is also transferable to $\Delta m(\alpha)$, with the only difference that in case of $\Delta\beta(\alpha)$ the limit will always be zero, independently of the CpGs, sample, and the tissue chosen, whereas in case of $\Delta m(\alpha)$ the limit depends on the CpG, sample, and tissue under consideration.

The convergence results for $\Delta\beta(\alpha)$ and $\Delta m(\alpha)$ imply that the percentage of CpGs satisfying the condition $\Delta\beta(\alpha) > 0$ and $\Delta m(\alpha) > 0$ for a given sample, respectively, approaches a positive constant as α increases. Standard computations verify this limit value to be just the percentage of CpGs satisfying $M_{BS} > M_{oxBS}$ for a given sample; see S2 Appendix for details.

Inspired by the convergence results obtained for the measures $\Delta\beta(\alpha)$ and $\Delta m(\alpha)$, we next propose

$$\Delta m^{\infty} = \log_2 \frac{M_{BS}}{M_{oxBS}} \left(= \lim_{\alpha \uparrow \infty} \Delta m(\alpha)\right) \tag{10}$$

as the first alternative 5hmC measure that can be used for the detection of hydroxymethylated CpGs. Note that Δm^{∞} is well-defined for all CpGs satisfying $M_{BS} > 0$ and $M_{oxBS} > 0$ simultaneously.

The main advantage of the measure Δm^{∞} in comparison to the measures $\Delta\beta(\alpha)$ and $\Delta m(\alpha)$ is its complete independence of the correction term α; this fact makes Δm^{∞} more robust for application in the screening step. Furthermore, the sign of Δm^{∞} has a very intuitive interpretation. Indeed, we get $\Delta m^{\infty} > 0$ if $M_{BS} > M_{oxBS}$ holds, i.e., if the global methylated intensity M_{BS} exceeds the "adjusted" methylated intensity M_{oxBS}. In all other cases we will have $\Delta m^{\infty} \leq 0$; for instance, $\Delta m^{\infty} = 0$ implies $M_{BS} = M_{oxBS}$, which can intuitively be interpreted as "no substantial 5hmC level observed".

In the context of our screening procedure, the most crucial question concerns a relation between the subsets of CpGs satisfying $\Delta m(\alpha) > 0$ and $\Delta m^{\infty} > 0$, respectively. To answer this question in a formal way, we divided the set of all CpGs with $\Delta m(\alpha) > 0$ in several disjoint subsets, and showed that, for a given sample and increasing α, the union of these subsets converges to the subset of CpGs satisfying $M_{BS} > M_{oxBS}$; see S2 Appendix for more details.

Due to its definition, Δm^{∞} does not take into account the unmethylated intensities U_{BS} and U_{oxBS}. This may become an issue even if the role of these intensities in the detection of hydroxymethylated CpGs has not been clarified yet. We address this issue by proposing another measure for selecting CpGs with a substantial level of hydroxymethylation, namely,

$$\Delta h = 1 - \frac{M_{oxBS} + U_{oxBS}}{M_{BS} + U_{BS}}. \tag{11}$$

In (11), $M_{BS} + U_{BS}$ is the global intensity obtained from the *BS* procedure and $M_{oxBS} + U_{oxBS}$ is the global intensity derived by means of the *oxBS* procedure.

For CpGs with $M_{BS} + U_{BS}$ exceeding $M_{oxBS} + U_{oxBS}$, i.e., for those CpGs which can be intuitively interpreted as exhibiting a substantial level of hydroxymethylation, the measure Δh must range between 0 and 1. In particular, the values of Δh close to zero correspond to $M_{BS} + U_{BS}$ being approximately equal to $M_{oxBS} + U_{oxBS}$ and thus the global 5hmC level being (almost) negligible. On the other hand, for Δh approximately equal to one we deduce that $M_{oxBS} + U_{oxBS}$ must be substantially smaller than $M_{BS} + U_{BS}$ and thus the global 5hmC level has to be high. Altogether, larger values of Δh correspond to larger proportions of the global 5hmC levels and we can interpret Δh as *the proportion of 5hmC in the global methylation*.

Our intuition in the interpretation of the values of Δh is based on the assumption that a substantial 5hmC level is associated with a substantial decrease in the overall intensities $M + U$, with $M_{BS} + U_{BS} > M_{oxBS} + U_{oxBS}$ for a given CpG site. Such interpretation is induced by the fact that, in contrast to the methylation process, a role of the unmethylated intensities U in the hydroxymethylation process is unclear. Thus, negative values of Δh are currently treated as a measurement error. Note that in (11) one has to assume that $M_{BS} + U_{BS}$ is different from zero; in other words, all CpGs with $M_{BS} + U_{BS}$ equal to zero have to be excluded from the analysis as exhibiting measurement error.

In view of the screening procedure, we also analyzed whether positive values of the measure Δh lead to positivity of other 5hmC measures introduced above, and vice versa. As (11) implies, the inequality $\Delta h > 0$ holds for

$$M_{BS} + U_{BS} > M_{oxBS} + U_{oxBS}. \tag{12}$$

However, the latter inequality is not sufficient to make a statement about the sign of the measures $\Delta\beta(\alpha)$ and Δm^{∞}, so that additional assumptions are needed; see S2 Appendix for details.

Altogether, our discussion indicates that the application of Δh for the detection of hydroxymethylated CpGs can be of advantage, since this 5hmC measure overcomes the limitation of both 5hmC measures considered earlier. In particular, this measure does not depend on the

choice of the correction term α, has an intuitive interpretation of its outcomes in terms of the observed 5hmC level, and can be computed directly from measured array data.

Materials and methods

Numerical analyses of the resemblance of $\Delta\beta(\alpha)$, Δm^{∞} and Δh

In the previous sections we considered three 5hmC measures, $\Delta\beta(\alpha)$, Δm^{∞} and Δh, as possible tools for the classification of CpGs into hydroxymethylated and those which do not exhibit a substantial level of hydroxymethylation. To estimate a possible classification error, one would usually compare each of these 5hmC measures with a certain gold standard. However, no gold standard is available in our case, since even the actual meaning of the formulations "a substantial 5hmC level observed" or "no substantial 5hmC level observed" in terms of measured methylated and unmethylated intensities M and U is unclear so far. One of possible ways to evaluate the accuracy of $\Delta\beta(\alpha)$, Δm^{∞} and Δh in the *absence of a gold standard*, as proposed in this section, is to describe this accuracy in terms of relative sensitivities and specificities of these measures with respect to each other. On the other hand, the resemblance of the considered 5hmC measures with respect to each other can also be addressed by means of a similarity analysis.

Numerical analyses of the present section were motivated by the discussions presented in [24, 43–48].

Study cohort, 5hmC isolation, data preprocessing

All analyses were performed on 38 paired samples, with both (colorectal) cancer and normal tissue available for each sample. All 38 patients were enrolled in the ongoing population-based case-control study DACHS (Darmkrebs: Chancen der Verhütung durch Screening, http://dachs.dkfz.org/dachs/), extensively described in [49]. Data collecting and patient recruitment procedures as well as the processes of DNA isolation and methylation profiling using the Infinium HumanMethylation450 BeadChip array (Illumina) are similar to those described in [50]. All data are publicly available at https://zenodo.org/record/2639285#.XLYzNKZS_XE.

All data analyses were performed using the computational environment R, V.3.5.2 (http://www.r-project.org/). Raw data signals from each of the *BS*- and *oxBS*- converted samples were preprocessed using the R/Bioconductor *minfi*-package [51]. In particular, the procedure *preprocessRaw* from that package was applied in order to convert the red/green channel for an Illumina methylation array into methylation signal.

Results

Prevalence of positive results

We applied all three considered 5hmC measures to both healthy and cancer tissue and computed the percentage of CpGs satisfying $\Delta\beta(100) > 0$, $\Delta m^{\infty} > 0$ and $\Delta h > 0$ for each given sample; Fig 6 illustrates the obtained results. Note that such *prevalence of positive results* is crucial in the screening procedure and represents the most intuitive approach for the comparison of any two 5hmC measures. Dependence of the prevalence of positive results of the measure $\Delta\beta(\alpha)$ on the choice of α is discussed in S1 Appendix.

Further, we adopted the statement in [24] on a reduction of 5hmC levels in cancer tissue to the prevalence of positive results, by expecting this prevalence to be higher in healthy tissue compared to cancer one. In a sample-wise analysis, this anticipation was indeed confirmed for the 5hmC measure $\Delta\beta(100)$, but not for the measures Δm^{∞} and Δh.

The same analysis, performed CpG-wise, i.e., with prevalence of positive results computed for each single CpG across all 38 samples that provides the *hydroxymethylation level* for each

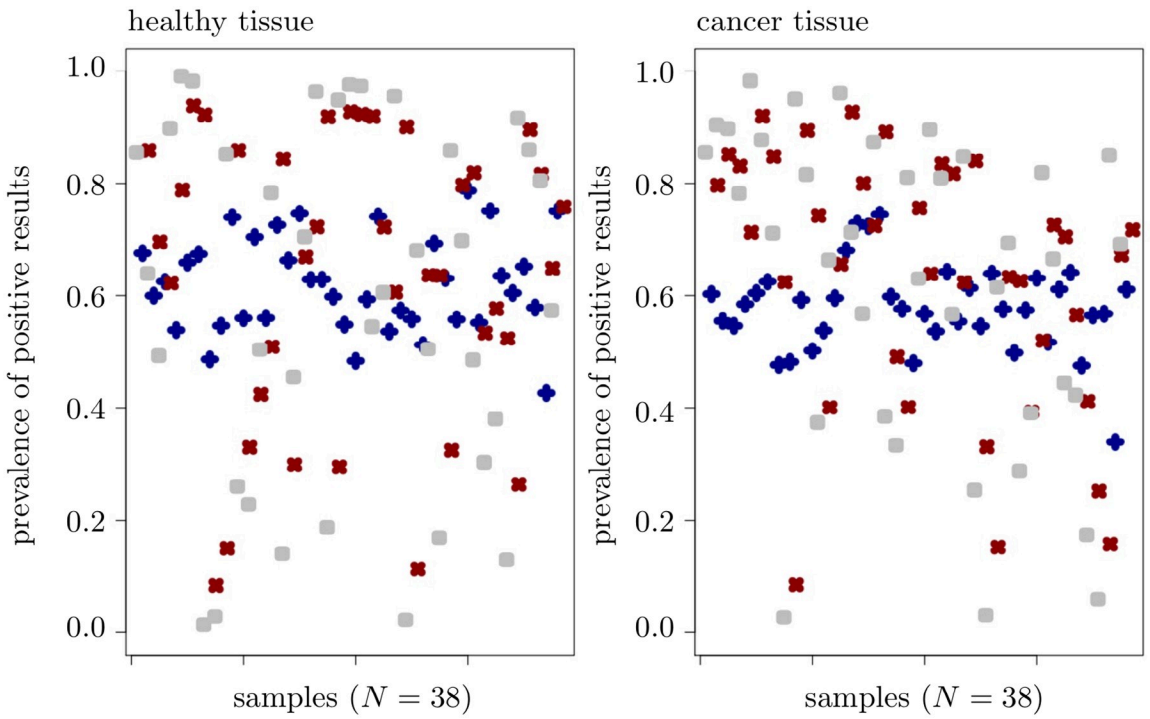

Fig 6. Sample-wise prevalence of positive results. The dark blue dots correspond to the percentage of CpGs with $\Delta\beta(100) > 0$, the dark red dots to $\Delta m^{\infty} > 0$ and the grey dots to $\Delta h > 0$.

https://doi.org/10.1371/journal.pone.0218103.g006

given CpG, again showed a significant reduction in 5hmC levels as obtained on cancer tissue, in particular for the measures $\Delta\beta(100)$ and Δm^{∞}. Contrary to our expectations, for the measure Δh, prevalence of positive results was significantly lower in healthy tissue compared to cancer one.

Next, we compared prevalences of positive results of any two 5hmC measure on a given tissue, in order to investigate the *conservativeness* of these measures when screening for hydroxymethylated CpGs. This analysis, performed sample-wise, resulted in the 5hmC measure Δm^{∞} being less conservative than Δh on healthy tissue and less conservative than $\Delta\beta(100)$ on cancer tissue; see Fig 6 for an illustration. The same analysis, performed CpG-wise, determined Δm^{∞} as being the least conservative 5hmC measure, on both considered tissues. Further, on healthy tissue $\Delta\beta(100)$ appeared to be less conservative than Δh, whereas on cancer tissue, Δh was less conservative than $\Delta\beta(100)$. This result can be interpreted as an evidence of the *tissue effect* [41, 52].

We also analyzed the *joint prevalence of positive results* defined as the percentage of CpGs with any two 5hmC measures being positive; such joint prevalence characterizes the *agreement* between any two 5hmC measures in the context of the screening step. Sample-wise analysis did not reveal any significant differences in these joint prevalences as calculated on healthy and cancer tissue. On the other hand, the joint prevalence of the measures $\Delta\beta(100)$ and Δm^{∞} appeared to exceed the joint prevalence of $\Delta\beta(100)$ and Δh significantly, on both considered tissues. The same result, again for both tissues, holds for the joint prevalences of the measures Δm^{∞} and Δh as well as of the measures $\Delta\beta(100)$ and Δh. Finally, on cancer tissue, the joint prevalence of the measures $\Delta\beta(100)$ and Δm^{∞} significantly exceeded the joint prevalence of the measures Δm^{∞} and Δh. In total, we conclude that, in a sample-wise analysis performed on cancer tissue, the 5hmC measures $\Delta\beta(100)$ and Δm^{∞} demonstrate the strongest agreement, followed by agreement between the measures Δm^{∞} and Δh.

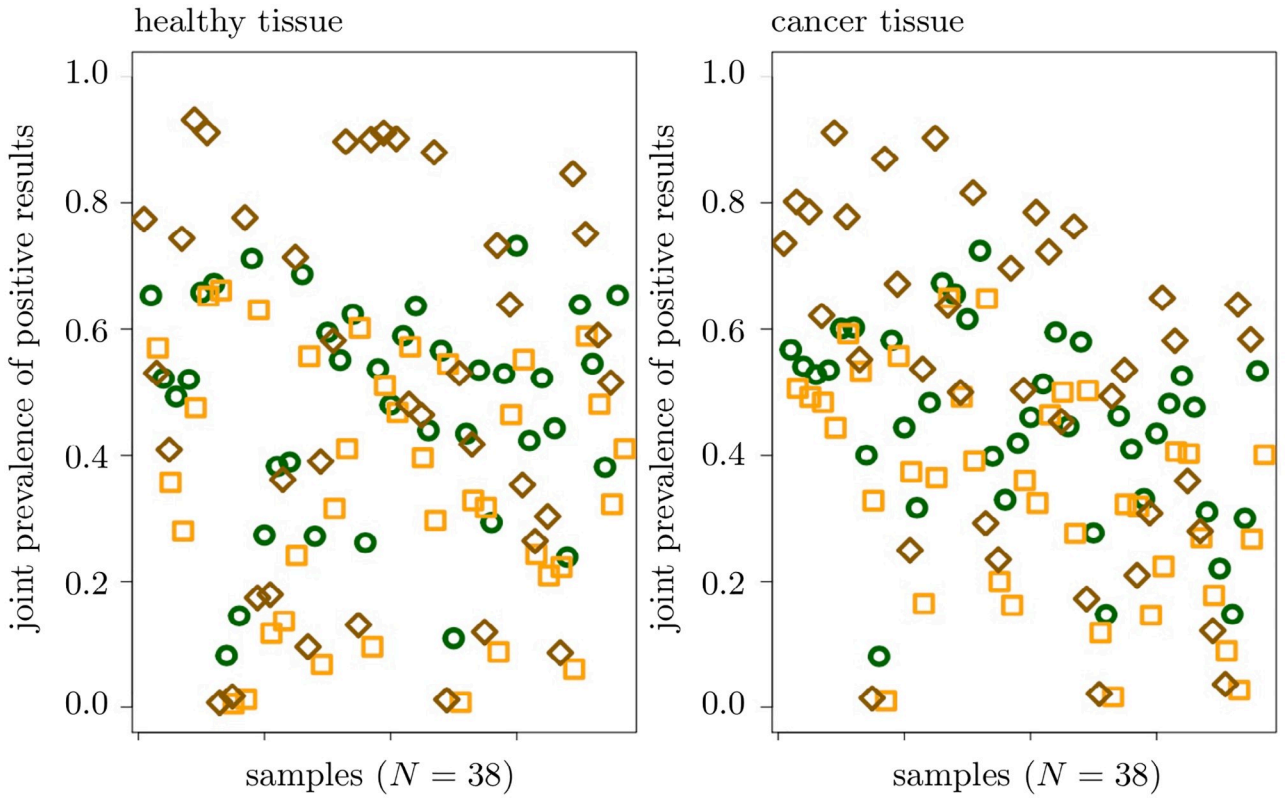

Fig 7. Sample-wise joint prevalence of positive results. Orange squares correspond to the values for $\Delta\beta(100)$ and Δh, dark green circles to the values for $\Delta\beta(100)$ and Δm^{∞} and brown squares to the values for Δh and Δm^{∞}.

https://doi.org/10.1371/journal.pone.0218103.g007

The same joint prevalence analysis, performed CpG-wise, revealed the joint prevalence of the measures $\Delta\beta(100)$ and Δm^{∞} on healthy tissue being significantly higher than the corresponding joint prevalence on cancer tissue; similar result is true for the joint prevalence of the measures $\Delta\beta(100)$ and Δh. As in case of a sample-wise analysis, the joint prevalence of the measures $\Delta\beta(100)$ and Δm^{∞} significantly exceeded the joint prevalence of $\Delta\beta(100)$ and Δh, both on healthy and cancer tissue; the same relation is true for the joint prevalences of the measures Δm^{∞} and Δh and of the measures $\Delta\beta(100)$ and Δh. On the other hand, in contrast to the results of the sample-wise analysis above, the joint prevalence of the measures $\Delta\beta(100)$ and Δm^{∞} is significantly lower than the joint prevalence of the measures Δm^{∞} and Δh, both on healthy and cancer tissue. Altogether, the CpG-wise analysis showed the highest agreement between the measures Δm^{∞} and Δh, followed by the agreement between the measures $\Delta\beta(100)$ and Δm^{∞}; the 5hmC measures $\Delta\beta(100)$ and Δh demonstrated the lowest pairwise agreement, consistent with the results of the sample-wise analysis. Sample-wise joint prevalence of positive results is visualized in Fig 7. Joint agreement between all three 5hmC measures is illustrated in Fig 8; for more results see also Figs A and B in S3 Appendix.

To summarize the results of our discussion above, we state that the 5hmC measure $\Delta\beta(100)$ demonstrates a higher agreement with Δm^{∞} than with Δh. Moreover, the agreement between the measures Δm^{∞} and Δh exceeds the agreement between $\Delta\beta(100)$ and Δh.

Similarity analyses

In order to address the resemblance of the proposed 5hmC measures without making any statement about their performance, *similarity analyses* can also be applied; the main tool of

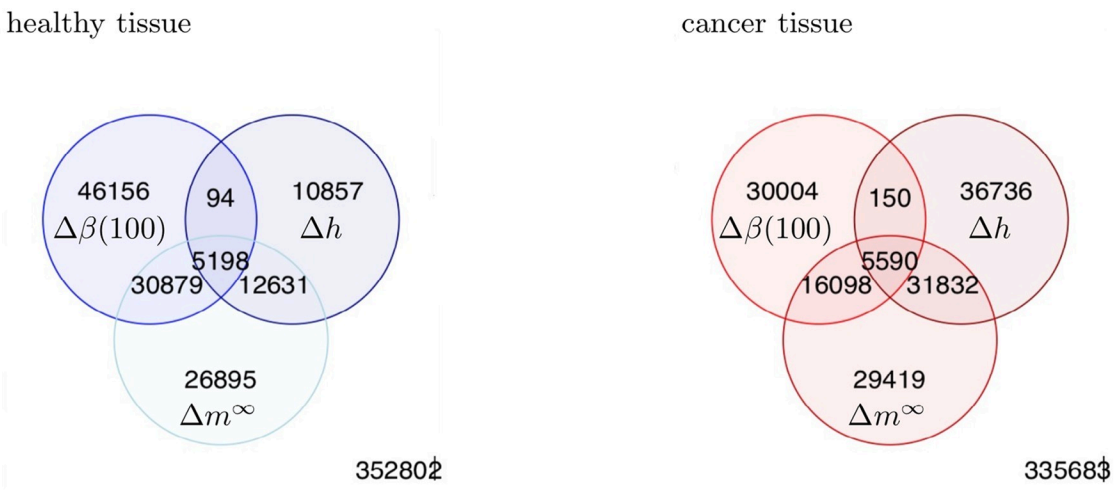

Fig 8. The number of substantially hydroxymethylated CpGs as identified by all three 5hmC measures. The number of substantially hydroxymethylated CpGs as identified by all three 5hmC measures, on healthy (the left-hand panel) and cancer (the right-hand panels) tissues and across all 38 samples. A CpG site is considered to be substantially hydroxymethylated under a given 5hmC measure x, if at least 75% of all values of x computed for this CpG and across all 38 samples are positive.

https://doi.org/10.1371/journal.pone.0218103.g008

such analyses is a *similarity coefficient*. There is a variety of similarity coefficients proposed in literature. For an overview see, e.g., [53–56] and references therein.

In order to quantify the *pairwise similarity* of the proposed 5hmC measures $\Delta\beta(\alpha)$, Δm^{∞}, and Δh in the context of the screening step, we first considered the similarity coefficient $\mathbb{S}$, also known as *the simple matching coefficient* [53, 54]. In particular, for a given CpG and two given 5hmC measures x_1 and x_2 we rewrite this similarity coefficient as

$$\mathbb{S}(x_1, x_2) = \frac{1}{n}\left(\sum_{i=1}^{n} I_{\{x_1^i>0\}} I_{\{x_2^i>0\}} + \sum_{i=1}^{n} I_{\{x_1^i\leq 0\}} I_{\{x_2^i\leq 0\}}\right). \tag{13}$$

Here n is the number of samples under consideration, $I_{\{x>0\}}$ is the indicator function, with $I_{\{x>0\}} = 1$ for $x > 0$ and $I_{\{x>0\}} = 0$ otherwise, and x_j^i is the value of the measure $x_j (j = 1, 2)$ in the ith CpG. Clearly, the similarity coefficient $\mathbb{S}$ in (13) ranges between 0 and 1, with 1 corresponding to *complete similarity* and 0 to *complete dissimilarity* between the considered two measures x_1 and x_2. Moreover, the similarity coefficient $\mathbb{S}$ represents an extension of the prevalence of positive results introduced earlier, since it considers not only the CpG sites that were flagged as hydroxymethylated but also those CpG sites that were identified as non-hydroxymethylated by two considered 5hmC measures.

While performing the similarity analysis for each given sample, we could not state any significant difference in the values of $\mathbb{S}$ as computed on healthy and cancer tissue. Further, the 5hmC measures Δm^{∞} and Δh appears to be the most similar, whereas the measures Δh and $\Delta\beta$ (100) are the least similar, both on healthy and cancer tissue. Finally, the 5hmC measure $\Delta\beta$ (100) is less similar to Δh than to Δm^{∞}, both on healthy and cancer tissue. All these results are visualized in Fig 9.

To describe the distribution of $\mathbb{S}$ for any two given 5hmC measures, we adapted the ideas of [56, 57] and calculated the expected value of this similarity coefficient. The results of those calculations are presented in the left-hand panel of Table 1. Due to that table, on cancer tissue the measures Δm^{∞} and Δh are again the most similar 5hmC measures; further, $\Delta\beta(100)$ and Δh are the least similar to each other, both on healthy and cancer tissue.

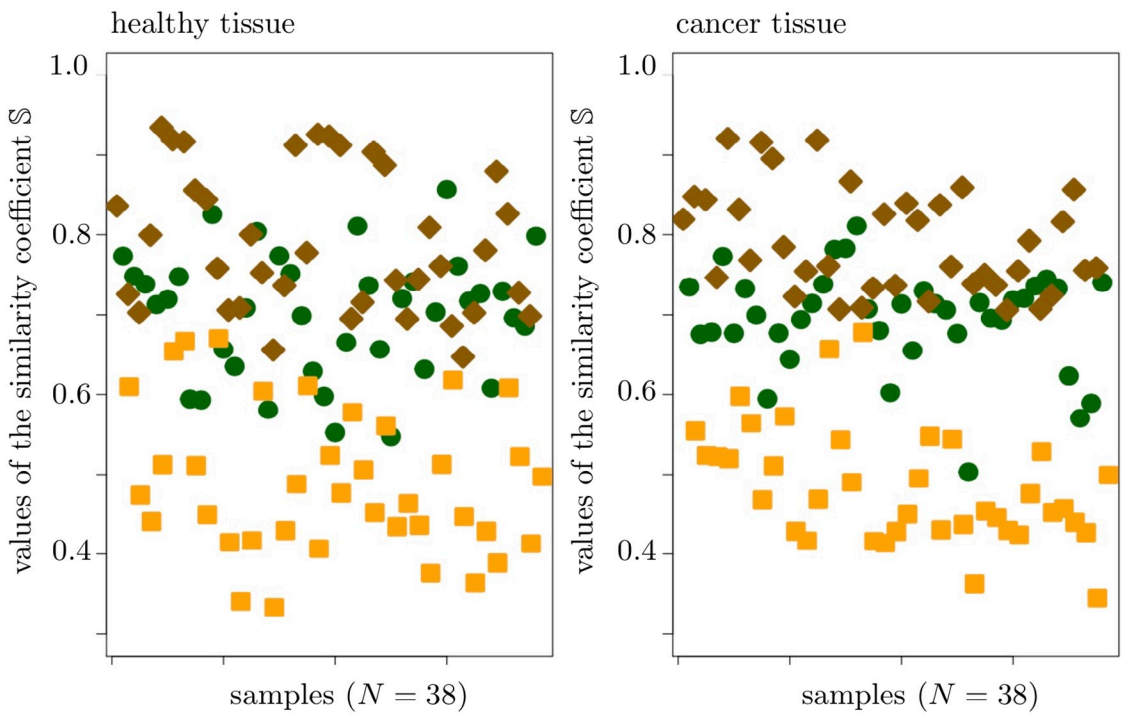

Fig 9. Pairwise similarity of the 5hmC measures $\Delta\beta(100)$, Δm^{∞} and Δh, in terms of the similarity coefficient $\mathbb{S}$. Orange rectangles correspond to the values of $\mathbb{S}(\Delta\beta(100), \Delta h)$, dark green dots to the values of $\mathbb{S}(\Delta\beta(100), \Delta m^{\infty})$ and brown rectangles to the values of $\mathbb{S}(\Delta h, \Delta m^{\infty})$.

https://doi.org/10.1371/journal.pone.0218103.g009

The similarity coefficient $\mathbb{S}$ in (13) exhibits a number of advantages such as simple applicability and intuitive interpretation of the obtained values. However, there are also some issues related to this coefficient. One of these issues arises in situations with two 5hmC measures x_1 and x_2 characterized by

$$\sum_{i=1}^{n} I_{\{x_1^i>0\}} I_{\{x_2^i>0\}} = 0. \tag{14}$$

For such 5hmC measures, which should actually be considered as completely dissimilar in the context of 5hmC detection, there is still a real possibility to get a positive value of the coefficient $\mathbb{S}$ as

$$\mathbb{S}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_1^i\le 0\}} I_{\{x_2^i\le 0\}} \tag{15}$$

Table 1. Expected values of the similarity coefficients $\mathbb{S}$ and $\mathbb{S}_H$.

	$\mathbb{E}[\mathbb{S}(.,.)]$		$\mathbb{E}[\mathbb{S}_H(.,.)]$	
5hmC measures	**healthy tissue**	**cancer tissue**	**healthy tissue**	**cancer tissue**
$\Delta\beta(100)$, Δh	0.7946	0.7897	0.5892	0.5794
$\Delta\beta(100)$, Δm^{∞}	0.8052	0.8016	0.6104	0.6032
Δm^{∞}, Δh	0.8021	0.8060	0.6042	0.6120

Expected values of the similarity coefficients $\mathbb{S}$ and $\mathbb{S}_H$ as applied for the pairwise comparison of the 5hmC measures $\Delta\beta(100)$, Δm^{∞} and Δh.

https://doi.org/10.1371/journal.pone.0218103.t001

which may indeed become misleading in the context of the screening step. This situation will even deteriorate for $\sum_{i=1}^{n} I_{\{x_1^i \le 0\}} I_{\{x_2^i \le 0\}} \to n$.

To mitigate this issue, we consider the *similarity coefficient of Hamann*, $\mathbb{S}_H$, defined as

$$\begin{aligned} \mathbb{S}_H(x_1, x_2) \quad &= \frac{1}{n}\left(\sum_{i=1}^{n} I_{\{x_1^i > 0\}} I_{\{x_2^i > 0\}} + \sum_{i=1}^{n} I_{\{x_1^i \le 0\}} I_{\{x_2^i \le 0\}} \right. \\ &\left. - \sum_{i=1}^{n} I_{\{x_1^i > 0\}} I_{\{x_2^i \le 0\}} - \sum_{i=1}^{n} I_{\{x_1^i \le 0\}} I_{\{x_2^i > 0\}} \right). \end{aligned} \tag{16}$$

Clearly, $\mathbb{S}_H$ is just a transformation of the simple matching coefficient $\mathbb{S}$ [53] that incorporates a correction for possible mismatches between the considered 5hmC measures x_1 and x_2. While ranging in the interval $[-1, 1]$, $\mathbb{S}_H(x_1, x_2) = -1$ can be interpreted as *complete dissimilarity* and $\mathbb{S}_H(x_1, x_2) = 1$ as *complete similarity* between x_1 and x_2. Further, due to (13) and (16), $\mathbb{S}_H(x_1, x_2) \le \mathbb{S}(x_1, x_2)$ for any two measures x_1 and x_2.

As in case with $\mathbb{S}$, we calculated the expected value of $\mathbb{S}_H$ for any two given 5hmC measures; the results are presented in the right-hand panel of Table 1. As expected, this table shows the similarity coefficient $\mathbb{S}_H$ confirming the results obtained under $\mathbb{S}$, e.g., with the 5hmC measures Δm^{∞} and Δh being most similar to each other on cancer tissue.

Altogether, we state that among three considered 5hmC measures $\Delta\beta(\alpha)$, Δm^{∞} and Δh, the measures Δm^{∞} and Δh appear to be most similar to each other on cancer tissue, both in terms of $\mathbb{S}$ and $\mathbb{S}_H$. Further, as in case of the prevalence of positive results analysis, the measure Δm^{∞} is more similar to $\Delta\beta(\alpha)$ than the measure Δh is.

Relative accuracy analyses

A different aproach for addressing the pairwise resemblance of the proposed 5hmC measures is to consider their *relative* sensitivities SE_r, specificities SP_r and false discovery rates FDR_r. Here, for any two 5hmC measures x_1 and x_2, we set $SE_r(x_1 \,|\, x_2) = \mathbb{P}(x_1 > 0 \,|\, x_2 > 0)$ as the relative sensitivity of x_1 with respect to x_2, $SP_r(x_1 \,|\, x_2) = \mathbb{P}(x_1 \le 0 \,|\, x_2 \le 0)$ as the relative specificity of x_1 with respect to x_2 and

$$FDR_r(x_1 \,|\, x_2) = 1 - \frac{\mathbb{P}(x_1 > 0, x_2 > 0)}{\mathbb{P}(x_1 > 0)} = 1 - SE_r(x_2 \,|\, x_1) \tag{17}$$

as the relative false discovery rate. The quantities SE_r and SP_r are also known as *co-positivities* and *co-negativities*, respectively [58, 59].

We started our data analyses on relative accuracies by checking for a significant difference in relative sensitivities as computed on healthy and cancer tissue. In a sample-wise analysis, such difference was observed for the relative sensitivities $SE_r(\Delta\beta(100)|\Delta h)$ and $SE_r(\Delta\beta(100)|\Delta m^{\infty})$, with the relative sensitivity on healthy tissue exceeding the corresponding relative sensitivity on cancer tissue. The same analysis, performed CpG-wise, showed all relative sensitivities differentiating significantly between healthy and cancer tissue.

Further, in a sample-wise analysis, performed on healthy tissue, the 5hmC measure Δm^{∞} demonstrated a higher sensitivity with respect to Δh than Δh did with respect to Δm^{∞}. This is consistent with our results on prevalence of positive results, with the measure Δm^{∞} being less conservative than Δh on healthy tissue. Further, there was a trend for a significant increase in the relative sensitivity $SE_r(\Delta m^{\infty}|\Delta\beta(100))$ compared to the relative sensitivity $SE_r(\Delta\beta(100)|\Delta m^{\infty})$ on cancer tissue. This is also related to our result on prevalence of positive results, with the measure Δm^{∞} being less conservative than $\Delta\beta(100)$ on cancer tissue.

A CpG-wise analysis of relative sensitivities revealed, the 5hmC measure $\Delta\beta(100)$ showing a lower sensitivity with respect to the measure Δh than the other way around, on cancer tissue. This result changed to the opposite on healthy tissue. Further, the measure Δm^{∞} showed a higher sensitivity with respect to the measure $\Delta\beta(100)$ than $\Delta\beta(100)$ did with respect to Δm^{∞}, both on healthy and cancer tissue. Analogous result was true for the measures Δm^{∞} and Δh, with $SE_r(\Delta m^{\infty}|\Delta h)$ exceeding $SE_r(\Delta h|\Delta m^{\infty})$, both on healthy and cancer tissue.

While analyzed sample-wise for its relative specificity, the measure $\Delta\beta(100)$ demonstrated a significantly lower specificity with respect to Δh on healthy tissue than on cancer tissue; similar result holds for relative specificity of the measure Δm^{∞} with respect to the measure Δh. The same analysis, performed CpG-wise, showed all relative specificities differentiating significantly between healthy and cancer tissue. Further, the 5hmC measure $\Delta\beta(100)$ demonstrated a higher specificity with respect to the measure Δm^{∞} than Δm^{∞} did with respect to $\Delta\beta(100)$, both on healthy and cancer tissue, with the difference being more substantial on cancer tissue. This is again in correspondence with the measure Δm^{∞} being less conservative than $\Delta\beta(100)$, in particular on cancer tissue.

In a CpG-wise analysis, on healthy tissue the measure $\Delta\beta(100)$ demonstrated a significantly lower specificity with respect to the measure Δh than Δh did with respect to $\Delta\beta(100)$; this result changes to the opposite while considering the same relative specificities on cancer tissue. Further, the measure Δm^{∞} showed a lower specificity with respect to $\Delta\beta(100)$ than $\Delta\beta(100)$ did with respect to Δm^{∞}, on both considered tissues; similar result is true for the measures Δm^{∞} and Δh.

Due to its definition, the results on relative false discovery rates FDR_r can be immediately derived from the corresponding results on SE_r. For instance, one can show that the measure Δh has a higher false discovery rate with respect to $\Delta\beta(100)$ than Δm^{∞}, both on healthy and cancer tissue.

We also computed expected relative sensitivities, specificities and false discovery rates of each 5hmC measure with respect to two others; the results are presented in Tables 2–4 below.

Altogether, due to our relative accuracy analyses, the measure Δm^{∞} again demonstrates more resemblance with $\Delta\beta(100)$ than the measure Δh, both on healthy and cancer tissue.

Comparison of $\Delta\beta(\alpha)$, Δh and Δm^{∞} to the oxBS-MLE and OxyBS procedures in the context of a screening step

When detecting CpGs with a substantial 5hmC level, one may compare the results provided by each of the considered three 5hmC measures $\Delta\beta(\alpha)$, Δh and Δm^{∞} with those derived from the oxBS-MLE and OxyBS procedures introduced in [44, 47]. When applied in a screening step, both oxBS-MLE and OxyBS procedures will flag the same cytosines as being

Table 2. Expected relative sensitivities $\mathbb{E}[SE_r(x_1 \mid x_2)]$.

	healthy tissue			cancer tissue		
$\mathbb{E}[SE_r]$	Δm^{∞}	$\Delta\beta(100)$	Δh	Δm^{∞}	$\Delta\beta(100)$	Δh
Δm^{∞}	1.0	0.7739	0.8624	1.0	0.7827	0.8383
$\Delta\beta(100)$	0.7456	1.0	0.5938	0.7133	1.0	0.5549
Δh	0.7940	0.5613	1.0	0.8199	0.5906	1.0

Expected relative sensitivities $\mathbb{E}[SE_r(x_1 \mid x_2)]$, computed for any two 5hmC measures $x_1, x_2 \in \{\Delta\beta(100), \Delta m^{\infty}, \Delta h\}$. The measure x_1 is in rows and x_2 is in columns; e.g., the value 0.7456 corresponds to the expected relative sensitivity $\mathbb{E}[SE_r(\Delta\beta(100) \mid \Delta m^{\infty})]$ and the value 0.7739 to the expected relative sensitivity $\mathbb{E}[SE_r(\Delta m^{\infty} \mid \Delta\beta(100))]$, both on healthy tissue. Since larger values in the table describe a higher relative sensitivity, our results in the table above indicate the measures Δh and Δm^{∞} demonstrating the highest resemblance. At the same time, the measures $\Delta\beta(100)$ and Δh appear to be least similar to each other.

https://doi.org/10.1371/journal.pone.0218103.t002

Table 3. Expected relative specificities $\mathbb{E}[SP_r(x_1 \mid x_2)]$.

	healthy tissue			cancer tissue		
$\mathbb{E}[SP_r]$	Δm^{∞}	$\Delta\beta(100)$	Δh	Δm^{∞}	$\Delta\beta(100)$	Δh
Δm^{∞}	1.0	0.5698	0.6844	1.0	0.5669	0.7112
$\Delta\beta(100)$	0.5982	1.0	0.3473	0.6455	1.0	0.3834
Δh	0.7890	0.3788	1.0	0.7422	0.3529	1.0

Expected relative specificities $\mathbb{E}[SP_r(x_1 \mid x_2)]$, computed for any two 5hmC measures $x_1, x_2 \in \{\Delta\beta(100), \Delta m^{\infty}, \Delta h\}$. The measure x_1 is in rows and x_2 is in columns; e.g., the value 0.5982 corresponds to the expected relative specificity $\mathbb{E}[SP_r(\Delta\beta(100) \mid \Delta m^{\infty})]$ and the value 0.5698 to the expected relative specificity $\mathbb{E}[SP_r(\Delta m^{\infty} \mid \Delta\beta(100))]$, both on healthy tissue. As in Table 2, larger values correspond to a higher relative specificity, and thus $\Delta\beta(100)$ and Δm^{∞} demonstrate a higher resemblance than $\Delta\beta(100)$ and Δh.

https://doi.org/10.1371/journal.pone.0218103.t003

hydroxymethylated as the 5hmC measure $\Delta\beta(0)$ will do. This results follows immediately from the problem formulations and the derivation of the MLEs as suggested by both procedures; see S4 Appendix for details. Thus, the comparison of the 5hmC measures Δm^{∞} and Δh with the oxBS-MLE and OxyBS procedures in detection of hydroxymethylated cytosines can be traced back to the comparison of these measures with the measure $\Delta\beta(0)$.

Conclusion

Presently, the measure most commonly used for the detection of hydroxymethylated CpGs is the measure $\Delta\beta(\alpha)$ and its derivatives as introduced in [31, 41, 42]. Well-established due to its easy computation and alleged intuitivity, this 5hmC measure nevertheless exhibits a number of limitations and has already been criticized due to its interpretation. This interpretation has meanwhile been questioned in [44], where the authors discussed the "naive" estimation of the 5hmC level via the difference of two β values as proposed in [31, 41] and introduced a model for describing the 5mC and 5hmC proportions by means of maximum likelihood estimation and beta-distributed random variables. Such modeling disallows negative proportions in particular; the corresponding model was also implemented in the *R*-package *OxyBS* [44].

In this paper, we performed a detailed analysis of $\Delta\beta(\alpha)$, both analytically and empirically, and discussed a number of limitations of $\Delta\beta(\alpha)$ which could make its practical applicability for screening of hydroxymethylated CpGs questionable. These limitations concern in particular the interpretation of $\Delta\beta(\alpha)$ and its robustness with respect to the choice of α.

Further, we proposed two alternative 5hmC measures which can be applied in the screening step. The first of these 5hmC measures is the measure Δm^{∞}. While intuitively interpretable and independent of the correction term α, this measure does not incorporates the unmethylated intensities U_{BS} and U_{oxBS}. Even though the role of these intensities in detection of the

Table 4. Expected relative false discovery rates $\mathbb{E}[FDR_r(x_1 \mid x_2)]$.

	healthy tissue			cancer tissue		
$\mathbb{E}[FDR_r]$	Δm^{∞}	$\Delta\beta(100)$	Δh	Δm^{∞}	$\Delta\beta(100)$	Δh
Δm^{∞}	0.0	0.2544	0.2060	0.0	0.2867	0.1801
$\Delta\beta(100)$	0.2261	0.0	0.4387	0.2173	0.0	0.4094
Δh	0.1376	0.4062	0.0	0.1617	0.4451	0.0

Expected relative false discovery rate $\mathbb{E}[FDR_r(x_1 \mid x_2)]$, computed for any two 5hmC measures $x_1, x_2 \in \{\Delta\beta(100), \Delta m^{\infty}, \Delta h\}$. The measure x_1 is in rows and x_2 is in columns; e.g., the value 0.2261 corresponds to the expected relative false discovery rate $\mathbb{E}[FDR_r(\Delta\beta(100) \mid \Delta m^{\infty})]$ and the value 0.2544 to $\mathbb{E}[FDR_r(\Delta m^{\infty} \mid \Delta\beta(100))]$, both on healthy tissue. Altogether, the measure Δm^{∞} demonstrates a lower false discovery rate with respect to $\Delta\beta(100)$ than Δh.

https://doi.org/10.1371/journal.pone.0218103.t004

5hmC levels has not been clarified yet, we took this fact into account and suggested the second alternative 5hmC measure, Δh. Due to its definition, this measure does not depend on the choice of α, has an intuitive interpretation in detecting hydroxymethylated CpGs, takes into account all intensities, and can be computed directly from the observed data.

The main challenge to be handled in our analysis referred to a mutual comparison of the considered 5hmC measures in the *absence of a gold standard*, as no biological or biochemical criterion for a CpG to be considered as "hydroxymethylated", e.g., in terms of methylated and non-methylated intensities M and U, is available so far. To overcome this challenge and to be able to address resemblance of the proposed 5hmC measures in the context of the screening step, we first analyzed the prevalences of positive results for each single 5hmC measure. Here, we first observed a decrease in this prevalence, while moving from healthy to cancer tissue, for the measures $\Delta\beta(\alpha)$ and Δm^{∞}. This result is also in accordance with the observation on a depletion of 5hmC levels in tumors compared to corresponding normal tissue as stated, e.g., in [24, 45, 60]. Moreover, the measure Δm^{∞} appears to be the measure with the largest prevalence of positive results, both on healthy and cancer tissue. In addition, data-based analysis of the joint prevalence of positive results revealed the strongest agreement between the measures Δm^{∞} and Δh, followed by the agreement between the measures $\Delta\beta(100)$ and Δm^{∞}; the 5hmC measures $\Delta\beta(100)$ and Δh demonstrated the lowest pairwise agreement. In other words, a stronger resemblance between the measures $\Delta\beta(\alpha)$ and Δm^{∞} than between the measures $\Delta\beta(\alpha)$ and Δh was observed so far. This result was also confirmed in the context of a similarity analysis as performed for a pairwise comparison of the proposed 5hmC measures.

In order to estimate relative accuracies of $\Delta\beta(100)$, Δm^{∞}, and Δh with respect to each other, we also used relative sensitivity and specificity analyses. As a result of those analyses, the measure Δm^{∞} demonstrated a higher sensitivity and a lower specificity with respect to $\Delta\beta(100)$ than vice versa; the same result holds for the measures Δm^{∞} and Δh. Moreover, we observed that the measure Δh has a higher false discovery rate with respect to $\Delta\beta(100)$ than Δm^{∞}. Altogether, we concluded, that, in the context of the screening step, the 5hmC measure Δm^{∞} exhibits more resemblance with the measure $\Delta\beta(\alpha)$ than Δh does and thus this measure would be the first choice if looking for a possible substitute for $\Delta\beta(\alpha)$ with another 5hmC measure in the screening procedure.

Our numerical analyses are based on raw data, with no normalization method applied. There are a variety of reasons for this. First, some of our results (such as the convergence result for $\Delta\beta(\alpha)$) were derived analytically and thus do not depend on the data used for their illustration. Second, there is no consistent normalization method to be applied when quantifying the 5hmC levels [42, 44]. Third, a possible impact of a particular normalization method on the results of the 5hmC classification is currently not obvious to us and can in fact be considered as a topic of future research.

Nevertheless, we did check our results on the data normalized by three different normalization methods, *funNorm*, *SWAN* and *Illumina*, as available in the *R*-package *minfi* [51]; for more details see S5 Appendix. As a consequence of such normalized data analyses, we do observe some differences to our results as obtained on raw data. However, there is no evidence that such differences have any biological meaning and are not just a product of the normalization method applied. For instance, in some cases we observe a reduction in the prevalence of positive results of a given 5hmC measure as calculated on normalized data compared to raw data. On the other hand, a reduction in the 5hmC levels on cancer tissue as observed in terms of the measure $\Delta\beta(100)$ is confirmed for all three normalized data sets as well. The same is true for the measure Δm^{∞} being less conservative than Δh on healthy tissue. Further, the measures Δm^{∞} and Δh are the ones that are most similar to each other (in terms of the similarity

coefficient $\mathbb{S}$) followed by the measures $\Delta\beta(100)$ and Δm^{∞}, both on raw and normalized data; this result holds both for healthy and cancer tissue.

There are also differences in results on detection of the hydroxymethylated CpGs provided by different normalization procedures. For instance, on cancer tissue, the measure $\Delta\beta(100)$ shows a significant reduction in the prevalence of positive results calculated on the *Illumina* data compared to the prevalence computed on the *funNorm* data. Further, both on healthy and cancer tissue, the measures $\Delta\beta(100)$ and Δm^{∞} demonstrate the strongest similarity (in terms of the similarity coefficient $\mathbb{S}$) on the *funNorm* normalized data, followed by the *SWAN* normalized data; the similarity between $\Delta\beta(100)$ and Δm^{∞} on the *Illumina* normalized data is the lowest one.

In the present paper we discussed the possible applicability of the considered 5hmC measures for detection of hydroxymethylated CpGs in the screening procedure. The immediate question arising in this context is the question about the applicability of these measures for the quantification of the observed 5hmC levels, similar to the applicability of β values used for quantification of the methylation levels. Even if the measure Δh appears to provide the most intuitive interpretation in contrast to the remaining two 5hmC measures, this question is still a topic of future research.

Supporting information

S1 Appendix. On the 5hmC measure $\Delta\beta(\alpha)$. Sign change of $\Delta\beta(\alpha)$, sample-wise convergence of the CpG sets satisfying $\{\Delta\beta(\alpha) > 0\}$ as α increases, the role of α in similarity analyses.
(PDF)

S2 Appendix. On the 5hmC measures $\Delta m(\alpha)$. Relation between the measures $\Delta m(\alpha)$ and $\Delta\beta(\alpha)$, the 5hmC measure $\Delta m(\alpha)$ as a function of α (monotonicity, convergence, sign change of $\Delta m(\alpha)$), relation between the subsets $\{\Delta m(\alpha) > 0\}$ and $\{\Delta m^{\infty} > 0\}$, relation between the subsets $\{\Delta h > 0\}$, $\{\Delta\beta(\alpha) > 0\}$ and $\{\Delta m^{\infty} > 0\}$.
(PDF)

S3 Appendix. On the resemblance of $\Delta\beta(\alpha)$, Δm^{∞} and Δh: Numerical results. Prevalence of positive results, joint prevalence of positive results, similarity analyses, relative accuracy analyses (relative sensitivity and specificity).
(PDF)

S4 Appendix. A comparison of the 5hmC measures $\Delta\beta(\alpha)$, Δh and Δm^{∞} with the results of the oxBS-MLE and OxyBS procedures.
(PDF)

S5 Appendix. A comparison of numerical analyses on raw and normalized data.
(PDF)

Acknowledgments

Support by the German Federal Ministry of Education and Research (01ER1505a, 01ER1505b) and the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, is gratefully acknowledged. Moreover, the authors thank both reviewers for many helpful and constructive comments on previous versions of the manuscript.

Author Contributions

Conceptualization: Alla Slynko, Axel Benner.

Formal analysis: Alla Slynko.

Funding acquisition: Axel Benner.

Investigation: Alla Slynko.

Methodology: Alla Slynko, Axel Benner.

Visualization: Alla Slynko.

Writing – original draft: Alla Slynko.

Writing – review & editing: Axel Benner.

References

1. Bansal A, Pinney SE. DNA methylation and its role in the pathogenesis of diabetes. Pediatric diabetes. 2017; 18(3):167–177. https://doi.org/10.1111/pedi.12521 PMID: 28401680
2. Kuasne H, de Syllos Cólus IM, Busso AF, Hernandez-Vargas H, Barros-Filho MC, Marchi FA, et al. Genome-wide methylation and transcriptome analysis in penile carcinoma: uncovering new molecular markers. Clinical epigenetics. 2015; 7(1):46. https://doi.org/10.1186/s13148-015-0082-4 PMID: 25908946
3. Kulis M, Esteller M. DNA methylation and cancer. In: Advances in genetics. vol. 70. Elsevier; 2010. p. 27–56.
4. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics. 2010; 11(3):191. https://doi.org/10.1038/nrg2732 PMID: 20125086
5. Lima S, Hernandez-Vargasl H, Hercegl Z. Epigenetic signatures in cancer: Implications for the control of cancer. Current opinion in molecular therapeutics. 2010; 12(3):316–324. PMID: 20521220
6. Pries LK, Gülöksüz S, Kenis G. DNA methylation in schizophrenia. In: Neuroepigenomics in Aging and Disease. Springer; 2017. p. 211–236.
7. Xu X, Gammon MD, Hernandez-Vargas H, Herceg Z, Wetmur JG, Teitelbaum SL, et al. DNA methylation in peripheral blood measured by LUMA is associated with breast cancer in a population-based study. The FASEB Journal. 2012; 26(6):2657–2666. https://doi.org/10.1096/fj.11-197251 PMID: 22371529
8. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011; 98(4):288–295. https://doi.org/10.1016/j.ygeno.2011.07.007 PMID: 21839163
9. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011; 3(6):771–784. https://doi.org/10.2217/epi.11.105 PMID: 22126295
10. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in bioinformatics. 2013; 15(6):929–941. https://doi.org/10.1093/bib/bbt054 PMID: 23990268
11. Fan S, Huang K, Ai R, Wang M, Wang W. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. Genomics. 2016; 107(4):132–137. https://doi.org/10.1016/j.ygeno.2016.02.005 PMID: 26921858
12. Li D, Xie Z, Le Pape M, Dye T. An evaluation of statistical methods for DNA methylation microarray data analysis. BMC bioinformatics. 2015; 16(1):217. https://doi.org/10.1186/s12859-015-0641-x PMID: 26156501
13. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics. 2010; 11 (1):587. https://doi.org/10.1186/1471-2105-11-587 PMID: 21118553
14. Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Genome research. 2013; 23(9):1541–1553. https://doi.org/10.1101/gr.152231.112 PMID: 23804401
15. Triche TJ, Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA methylation beadarrays. Nucleic acids research. 2013; 41(7):e90–e90. https://doi.org/10.1093/nar/gkt090
16. Godderis L, Schouteden C, Tabish A, Poels K, Hoet P, Baccarelli AA, et al. Global methylation and hydroxymethylation in DNA from blood and saliva in healthy volunteers. BioMed research international. 2015; 2015. https://doi.org/10.1155/2015/845041 PMID: 26090450

17. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012; 149(6):1368–1380. https://doi.org/10.1016/j.cell.2012.04.027 PMID: 22608086

18. Booth MJ, Ost TW, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nature protocols. 2013; 8(10):1841. https://doi.org/10.1038/nprot.2013.115 PMID: 24008380

19. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PloS one. 2010; 5(1):e8888. https://doi.org/10.1371/journal.pone.0008888 PMID: 20126651

20. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science. 2009; 324(5929):929–930. https://doi.org/10.1126/science.1169786 PMID: 19372393

21. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009; 324(5929):930–935. https://doi.org/10.1126/science.1170116 PMID: 19372391

22. Bachman M, Uribe-Lewis S, Yang X, Williams M, Murrell A, Balasubramanian S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nature chemistry. 2014; 6(12):1049. https://doi.org/10.1038/nchem.2064 PMID: 25411882

23. Ecsedi S, Rodríguez-Aguilera J, Hernandez-Vargas H. 5-Hydroxymethylcytosine (5hmC), or How to Identify Your Favorite Cell. Epigenomes. 2018; 2(1):3. https://doi.org/10.3390/epigenomes2010003

24. Ficz G, Gribben JG. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? Genomics. 2014; 104(5):352–357. https://doi.org/10.1016/j.ygeno.2014.08.017 PMID: 25179374

25. Fu S, Wu H, Zhang H, Lian CG, Lu Q. DNA methylation/hydroxymethylation in melanoma. Oncotarget. 2017; 8(44):78163. https://doi.org/10.18632/oncotarget.18293 PMID: 29100458

26. Hahn MA, Szabó PE, Pfeifer GP. 5-Hydroxymethylcytosine: a stable or transient DNA modification? Genomics. 2014; 104(5):314–323. https://doi.org/10.1016/j.ygeno.2014.08.015 PMID: 25181633

27. Hill PW, Amouroux R, Hajkova P. DNA demethylation, TET proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: an emerging complex story. Genomics. 2014; 104(5):324–333. https://doi.org/10.1016/j.ygeno.2014.08.012 PMID: 25173569

28. Jin SG, Kadam S, Pfeifer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic acids research. 2010; 38(11):e125–e125. https://doi.org/10.1093/nar/gkq223 PMID: 20371518

29. Kudo Y, Tateishi K, Yamamoto K, Yamamoto S, Asaoka Y, Ijichi H, et al. Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. Cancer science. 2012; 103(4):670–676. https://doi.org/10.1111/j.1349-7006.2012.02213.x PMID: 22320381

30. Laird A, Thomson JP, Harrison DJ, Meehan RR. 5-hydroxymethylcytosine profiling as an indicator of cellular state. Epigenomics. 2013; 5(6):655–669. https://doi.org/10.2217/epi.13.69 PMID: 24283880

31. Nazor KL, Boland MJ, Bibikova M, Klotzle B, Yu M, Glenn-Pratola VL, et al. Application of a low cost array-based technique—TAB-Array—for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. Genomics. 2014; 104(5):358–367. https://doi.org/10.1016/j.ygeno.2014.08.014 PMID: 25179373

32. Santiago M, Antunes C, Guedes M, Sousa N, Marques CJ. TET enzymes and DNA hydroxymethylation in neural development and function—how critical are they? Genomics. 2014; 104(5):334–340. https://doi.org/10.1016/j.ygeno.2014.08.018 PMID: 25200796

33. Severin PM, Zou X, Schulten K, Gaub HE. Effects of cytosine hydroxymethylation on DNA strand separation. Biophysical journal. 2013; 104(1):208–215. https://doi.org/10.1016/j.bpj.2012.11.013 PMID: 23332073

34. Shen L, Zhang Y. 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. Current opinion in cell biology. 2013; 25(3):289–296. https://doi.org/10.1016/j.ceb.2013.02.017 PMID: 23498661

35. Tellez-Plaza M, Tang Wy, Shang Y, Umans JG, Francesconi KA, Goessler W, et al. Association of global DNA methylation and global DNA hydroxymethylation with metals and other exposures in human blood DNA samples. Environmental health perspectives. 2014; 122(9):946–954. https://doi.org/10.1289/ehp.1306674 PMID: 24769358

36. Thomson JP, Meehan RR. The application of genome-wide 5-hydroxymethylcytosine studies in cancer research. Epigenomics. 2017; 9(1):77–91. https://doi.org/10.2217/epi-2016-0122 PMID: 27936926

37. Wang T, Pan Q, Lin L, Szulwach KE, Song CX, He C, et al. Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. Human molecular genetics. 2012; 21(26):5500–5510. https://doi.org/10.1093/hmg/dds394 PMID: 23042784

38. Wen L, Tang F. Genomic distribution and possible functions of DNA hydroxymethylation in the brain. Genomics. 2014; 104(5):341–346. https://doi.org/10.1016/j.ygeno.2014.08.020 PMID: 25205307

39. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012; 336(6083):934–937. https://doi.org/10.1126/science.1220671 PMID: 22539555

40. Cui L, Chung TH, Tan D, Sun X, Jia XY. JBP1-seq: a fast and efficient method for genome-wide profiling of 5hmC. Genomics. 2014; 104(5):368–375. https://doi.org/10.1016/j.ygeno.2014.08.023 PMID: 25218799

41. Stewart SK, Morris TJ, Guilhamon P, Bulstrode H, Bachman M, Balasubramanian S, et al. oxBS-450K: a method for analysing hydroxymethylation using 450K BeadChips. Methods. 2015; 72:9–15. https://doi.org/10.1016/j.ymeth.2014.08.009 PMID: 25175075

42. Field SF, Beraldi D, Bachman M, Stewart SK, Beck S, Balasubramanian S. Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). PLoS One. 2015; 10(2):e0118202. https://doi.org/10.1371/journal.pone.0118202 PMID: 25706862

43. Green BB, Houseman EA, Johnson KC, Guerin DJ, Armstrong DA, Christensen BC, et al. Hydroxymethylation is uniquely distributed within term placenta, and is associated with gene expression. The FASEB Journal. 2016; 30(8):2874–2884. https://doi.org/10.1096/fj.201600310R PMID: 27118675

44. Houseman EA, Johnson KC, Christensen BC. OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. Bioinformatics. 2016; 32 (16):2505–2507. https://doi.org/10.1093/bioinformatics/btw158 PMID: 27153596

45. Li M, Gao F, Xia Y, Tang Y, Zhao W, Jin C, et al. Filtrating colorectal cancer associated genes by integrated analyses of global DNA methylation and hydroxymethylation in cancer and normal tissue. Scientific reports. 2016; 6:31826. https://doi.org/10.1038/srep31826 PMID: 27546520

46. Uribe-Lewis S, Stark R, Carroll T, Dunning MJ, Bachman M, Ito Y, et al. 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. Genome biology. 2015; 16(1):69. https://doi.org/10.1186/s13059-015-0605-5 PMID: 25853800

47. Xu Z, Taylor JA, Leung YK, Ho SM, Niu L. oxBS-MLE: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. Bioinformatics. 2016; 32(23):3667–3669. https://doi.org/10.1093/bioinformatics/btw527 PMID: 27522082

48. Zhu Y, Lu H, Zhang D, Li M, Sun X, Wan L, et al. Integrated analyses of multi-omics reveal global patterns of methylation and hydroxymethylation and screen the tumor suppressive roles of HADHB in colorectal cancer. Clinical epigenetics. 2018; 10(1):30. https://doi.org/10.1186/s13148-018-0458-3 PMID: 29507648

49. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case–control study. Annals of internal medicine. 2011; 154 (1):22–30. https://doi.org/10.7326/0003-4819-154-1-201101040-00004 PMID: 21200035

50. Gündert M, Edelmann D, Benner A, Jansen L, Jia M, Walter V, et al. Genome-wide DNA methylation analysis reveals a prognostic classifier for non-metastatic colorectal cancer (ProMCol classifier). Gut. 2019; 68(1):101–110. https://doi.org/10.1136/gutjnl-2017-314711 PMID: 29101262

51. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–1369. https://doi.org/10.1093/bioinformatics/btu049 PMID: 24478339

52. Nestor CE, Ottaviano R, Reddington J, Sproul D, Reinhardt D, Dunican D, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome research. 2012; 22(3):467–477. https://doi.org/10.1101/gr.126417.111 PMID: 22106369

53. Baroni-Urbani C, Buser MW. Similarity of binary data. Systematic Zoology. 1976; 25(3):251–259. https://doi.org/10.2307/2412493

54. Cheetham AH, Hazel JE. Binary (presence-absence) similarity coefficients. Journal of Paleontology. 1969;p. 1130–1136.

55. Hubalek Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biological Reviews. 1982; 57(4):669–689. https://doi.org/10.1111/j.1469-185X.1982.tb00376.x

56. Snijders TA, Dormaar M, Van Schuur WH, Dijkman-Caes C, Driessen G. Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes. Journal of Classification. 1990; 7 (1):5–31. https://doi.org/10.1007/BF01889701

57. Goodall D. The distribution of the matching coefficient. Biometrics. 1967;p. 647–656. https://doi.org/10.2307/2528419 PMID: 6080202

58. Buck AA, Gart JJ, et al. Comparison of a Screening Test and a Reference Test in Epidemiologic Studies. I. Indices of Agreements and their Relation to Prevalence. American Journal of Epidemiology. 1966; 83(3):586–92. https://doi.org/10.1093/oxfordjournals.aje.a120609 PMID: 5932702

59. Buck A, Gart J, et al. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. American Journal of Epidemiology. 1966; 83 (3):593–602. https://doi.org/10.1093/oxfordjournals.aje.a120610 PMID: 5932703

60. Udali S, De Santis D, Ruzzenente A, Moruzzi S, Mazzi F, Beschin G, et al. DNA methylation and Hydroxymethylation in primary Colon Cancer and synchronous hepatic metastasis. Frontiers in genetics. 2018; 8:229. https://doi.org/10.3389/fgene.2017.00229 PMID: 29375619