Supplementary Methods for „ Computational analysis of genome-wide DNA-methylation during the differentiation of human embryonic stem cells along the endodermal lineage“ – Chavez et al., Genome Research 2010

1. MEDIPS package overview

The MEDIPS software was developed for analyzing data derived from methylated DNA immunoprecipitation (MeDIP) experiments (Weber et al. 2005) followed by sequencing (MeDIP-seq). Nevertheless, functionalities like the saturation analysis may be applied to other types of sequencing data (e.g. ChIP-Seq). MEDIPS addresses several aspects in the context of MeDIP-seq data analysis. These are:

- estimating the reproducibility for obtaining full genome methylation profiles with respect to the total number of given short reads and to the size of the reference genome,
- analyzing the coverage of genome wide DNA sequence patterns (e.g. CpGs) with the given set of sequence reads,
- calculating a CpG enrichment factor as a quality control for the immunoprecipitation and for a rough impression of the overall amount of enriched methylated CpGs,
- calculating genome wide MeDIP-seq signal densities at a user specified resolution,
- calculating genome wide sequence pattern densities (e.g. CpGs) at a user specified resolution,
- plotting of calibration plots as a data quality check and for a visual inspection of the dependency between local sequence pattern (e.g. CpG) densities and MeDIP-seq signals,
- normalization of MeDIP-seq data with respect to local sequence pattern (e.g. CpG) densities,
- summarized methylation values for genome wide windows of a specified length or for user supplied regions of interest (ROIs),
- identification of differentially methylated regions on raw or normalized data comparing two sets of MeDIP-seq data and with respect to background data derived from input experiments,
- exporting raw and normalized data for visualization in common genome browsers (e.g. the UCSC genome browser (Kuhn et al. 2009)).

The input to MEDIPS is the result of the sequence mapping. MEDIPS can be applied to any genome of interest. The only limitation to its use, are the available genomes within Bioconductors (Gentleman et al. 2004) *BSgenome (Pages)* package. For a detailed description of the MEDIPS package, please see the tutorial as provided together with the package.

2. Modelling of MeDIP-seq data

2.1 Genome vector

In order to calculate the genome-wide short read coverage, a targeted data resolution has to be determined. In principle, a short read coverage can be calculated for each base position. Because the resolution of MeDIP-seq data is restricted by the size of the sonicated DNA fragments after amplification and size selection (typically between 0.2-1kb), a bin size of 50bp is considered as a reasonable compromise on data resolution and computational costs. Moreover, short reads generated by modern-day sequencers do not represent the full DNA fragments but are of shorter length (e.g. 36bp). Therefore, the data is smoothed by extending each read to a length according to the estimated average length of sequenced DNA fragments (here 400 bp), either along the + or along the - direction, as specified by the short read dependent strand information. MEDIPS divides each chromosome into bins of size 50 bp and subsequently calculates the short read coverage on this resolution. In the following, the bin representation of the genome is called the *genome vector*.

2.2 Reads per million (rpm)

For each pre-defined genomic bin, the genome vector stores the number of provided overlapping extended short reads (these are the raw MeDIP-seq signals). Based on the total number of provided short reads (*n*), the raw MeDIP-seq signals can be transformed into a reads per million (*rpm*) format in order to assure that coverage profiles derived from different biological samples are comparable, although generated from differing amounts of short reads. Let x_{bin_i} be the raw MeDIP-seq signal of the genomic bin *i*, where $i = 1,...,m$ and *m* is the total number of genomic bins, then the *rpm* value of the genomic bin is simply defined as:

$$rpm_{bin_i} = \frac{x_{bin_i} \cdot 10^6}{n}$$

MEDIPS allows for exporting WIG files containing genome wide *rpm* values at a user-specified resolution (here 50 bp). By utilizing these WIG files, the *rpm* profiles of the processed biological sample can be immediately visualised using a suitable genome browser.

2.3 Quality controls

2.3.1 Saturation analysis

MeDIP-seq aims to reconstruct methylation profiles on the basis of local short read coverages. It is supposed that an insufficient number of short reads will not represent the true methylation profile. Only when a sufficient number of short reads is generated, the resulting genome vector will represent a saturated methylation profile. Therefore, the saturation analysis addresses the question, whether the number of available short reads is sufficient to generate a saturated and reproducible methylation profile of the reference genome.

The basic assumption of the saturation analysis is that only a sufficient number of short reads will result in a genome wide methylation profile which will be reproducible by another independent set of a similar number of short reads. The correlation of two independently generated genome vectors will increase when the total number of short-reads considered for the construction of each of the two genome vectors increases. It is supposed that the increase of correlation between two independently generated genome vectors will saturate as soon as the total number of considered short reads is increased to a level that is able to represent the analysed methylome in a saturated way. Obviously, the number of short reads that have to be generated for a sufficient sequencing depth depends on the size of the reference genome.

For the saturation analysis, the total set of available regions (n) is divided into two distinct random sets A and B of equal size. Both sets A and B are again divided into k random subsets of equal size:

$$A = a_1,...,a_k$$

$$B = b_1,...,b_k$$

The saturation analysis runs in k iterations. For each set A and B independently, the saturation analysis iteratively selects an increasing number of subsets and creates according genome vectors by using an arbitrary bin size (here 50bp) and by previously extending the short reads to a suitable length (here 400bp). In each iteration step, the resulting genome vectors for the subsets of A and B are compared using Pearson correlation. As the number of considered short reads increases during each iteration step, it is supposed that the resulting genome vectors become more

similar, a dependency that is expressed by an increased correlation. By storing the resulting correlation coefficients after each iteration step, the change of correlation during the *k* iteration steps can be visualized by plotting the number of considered reads against the resulting correlation coefficients. Such a plot allows for gaining an impression of the reproducibility of constructing a methylome with respect to the number of considered short reads and with respect to the size of the reference genome.

However, such a saturation analysis can be performed on two independent sets of short reads, only. Therefore, a true saturation analysis can only be calculated for half of the available short reads. Obviously, it is of interest to examine the reproducibility of the MeDIP-seq experiment for the total amount of available short reads. Therefore, the saturation analysis is followed by an estimated saturation analysis. For the estimated saturation analysis, the full set of given regions (*n*) is artificially doubled by considering each region twice. Afterwards, the described saturation analysis is performed on the artificially doubled set of regions. Because the artificially doubled set of short reads does not represent a true outcome of a MeDIP-seq experiment, the calculated correlations will overestimate the true reproducibility. It is assumed that the true correlation for the full set of available short reads will be between the results of the true and of the estimated saturation analysis. Methods that randomly select data entries can be processed several times in order to obtain more stable results. Therefore, the random partitioning of the short reads into the several subsets of *A* and *B* was repeated ten times and the results were averaged.

2.3.2 Coverage analysis

The coverage analysis addresses the question about the genome wide depth of sequence pattern (here CpG) coverage by an increasing number of integrated sequencing derived short reads. For this, all genomic coordinates of the sequence pattern of interest have to be identified. The MEDIPS package provides a function for identifying the genomic positions of arbitrary sequence patterns. In the following, it is expected that all genomic pattern positions are stored on a vector $P = p_1,...,p_i,...,p_m$ where *m* is the number of sequence patterns present in the reference genome. For the coverage analysis, the total set of available short reads (*A*) is divided into *k* random subsets of equal size:

$$A = a_1,...,a_k$$

The coverage analysis runs in *k* iterations. The coverage analysis iteratively selects an increasing number of subsets and tests how many pattern positions from *P* are covered by the available

regions. In addition, the coverage analysis counts how many p_i's are covered at least Q times, where $Q = q_1,...,q_l$ represents an arbitrary number of coverage depths to be tested. For example, the according function of the MEDIPS package tests by default how many CpGs are covered at least 1x, 2x, 3x, 4x, 5x, and 10x times (this is equivalent to the notation Q = 1, 2, 3, 4, 5, 10). The *k*-th iteration step of the coverage analysis shows the depth of sequence pattern coverages obtained with the full set of available short reads.

The advantage of the iterative approach is that the behaviour of pattern coverage can be examined with respect to an increasing number of considered short reads. For this, coverage curves can be generated by plotting the number of covered sequence patterns after each iteration step and for each level of Q against the number of considered short reads. The progressions of the resulting coverage curves indicate the state of saturation of the overall sequence pattern coverages. Because methods that randomly select data entries can be processed several times in order to obtain more stable results, the random partitioning of the short reads into the several subsets of A was repeated ten times and the results were averaged. As for calculating the genome vector and as done for the saturation analysis the length of the short reads were previously extended to 400bp.

2.3.3 CpG enrichment

As a third MeDIP-seq data quality control, the CpG enrichment approach examines how strong the genomic regions underlying the obtained short reads are enriched for CpGs compared to the frequency of CpGs present in the reference genome. For this, firstly the number of cytosines (*G.c*), the number of guanines (*G.g*), the number CpGs (*G.cg*), and the total number of bases (*m*) within the specified reference genome (here hg19) are counted. Subsequently, the relative frequency of CpGs and the observed/expected (Gardiner-Garden and Frommer 1987) ratio of CpGs as present in the reference genome are calculated as:

$$Genome.CpG_{rel.f} = \frac{G.cg}{m}$$

$$Genome.CpG_{obs/\exp} = \frac{G.cg \cdot m}{G.c \cdot G.g}$$

Additionally, the number of cytosines (*SR.c*), the number of guanines (*SR.g*), the number CpGs (*SR.cg*), and the total number of bases (*n*) are counted for the DNA sequences underlying the

given short reads. Subsequently, the relative frequency of CpGs and the observed/expected ratio of CpGs as present in the short reads specific DNA sequences are calculated accordingly:

$$SR.CpG_{rel.f} = \frac{SR.cg}{n}$$

$$SR.CpG_{obs/\exp} = \frac{SR.cg \cdot n}{SR.c \cdot SR.g}$$

The final enrichment values result by dividing the relative frequency of CpGs (or the observed/expected value, respectively) of the short reads by the relative frequency of CpGs (or the observed/expected value, respectively) of the reference genome:

$$enrich_{rel.f} = \frac{SR.CpG_{rel.f}}{Genome.CpG_{rel.f}}$$

$$enrich_{obs/\exp} = \frac{SR.CpG_{obs/\exp}}{Genome.CpG_{obs/\exp}}$$

For short reads derived from an INPUT experiment (that is sequencing of none-enriched DNA fragments), the enrichment values are expected to be close to 1. In contrast, short reads derived from MeDIP-seq experiments are expected to be enriched for CpG rich DNA sequences, a circumstance which will be indicated by increased enrichment scores.

2.4 MeDIP-seq data normalization

The idea of a MeDIP experiment is to identify cytosine methylation profiles of a sample of interest by immunocapturing methylated CpGs (mCpGs) using an mCpG specific antibody (Weber et al. 2005). However, it has been shown (Down et al. 2008; Pelizzola et al. 2008) that MeDIP signals scale with local densities of CpGs and are not necessarily influenced by mCpGs, only. Therefore, the need for MeDIP-seq data correction occurs through an unspecific binding of the utilized antibody to un-methylated CpGs, especially in genomic regions associated to elevated densities of un-methylated CpGs and low densities of mCpGs.

2.4.1 Coupling factors

Similar to other MeDIP normalization approaches (Down et al. 2008; Pelizzola et al. 2008), the presented method corrects for the unspecific antibody binding by incorporating local CpG densities into the MeDIP-seq derived signals. In order to integrate the information about CpG

densities into the following analysis, it is necessary to identify the genomic positions of all CpGs. This can be achieved by executing the *MEDIPS.getPosition()* function of the MEDIPS package. Following the valuable concept of coupling factors presented by Down et al. (Down et al. 2008), a *coupling vector* is calculated based on the received genomic positions of all CpGs. The coupling vector is of the same size as the predefined genome vector (here bin size of 50bp) but contains local CpG denisties (also called coupling factors) for each genomic bin, instead. For each predefined genomic bin at position *b*, the density of surrounding CpGs has to be calculated. For this, first a maximal distance (*d*) has to be defined. Only CpGs within the range of $[b-d, b, b+d]$ will contribute to the final local coupling factor at *b*. The optimized value for *d* will reflect the estimated size of the sonicated DNA fragments after amplification and size selection. This is because MeDIP-seq derived signals at position *b* are influenced by sequenced DNA fragments that overlap with position *b*. Immunoprecipitation of these DNA fragments can be caused by a methylated and antibody bound CpG located at any position of the DNA-fragment. The maximal distance of a CpG contributing to the signal at *b* is therefore the estimated average length of the sonicated DNA fragments (*d*).

There are several ways for calculating coupling factors for genomic bins. Let *c* be the chromosomal position of a CpG and as *b* is the chromosomal position of a genomic bin, $dist = |b-c|$ is the distance between the genomic bin and the CpG. A CpG will contribute to the coupling factor of a genomic bin at position *b*, if $dist \leq d$. The simplest way is to count the number of CpGs within the maximal distance *d* around a genomic bin at position *b* (*count* function). Another approach is to weight each CpG by its distance to the current genomic bin. CpGs farther away from the current genomic bin will receive smaller weights, whereas CpGs close to the genomic bin will receive higher weights. The upper panel in Figure 1 illustrates a genome vector generated by defining a bin size of 50bp. In addition, CpGs are given in a schematic way. The Figure illustrates that immuoprecipitated DNA fragments of an estimated average length greater than the pre-defined bin size can contribute to the signal of the genomic bin at position *b* (vertical red line). Moreover, the schematic distance function illustrates that CpGs close to position *b* will receive higher weights than CpGs located farther away. There are several possible ways for defining weighting functions. In the context of this thesis, the following weighting functions were evaluated: *count, linear, exp (Pelizzola et al. 2008), log (Pelizzola et al. 2008),* and *custom (Down et al. 2008).* The images at the bottom of Figure 1 show the progression of these weighting functions by defining a maximal distance d = 700.

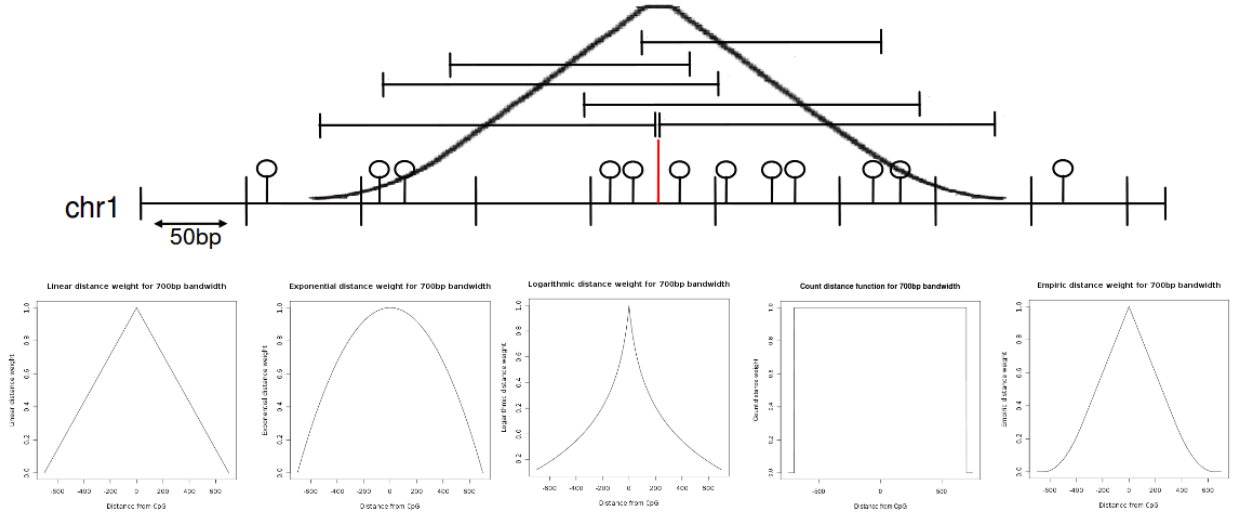

Figure 1: Calculation of coupling factors. The upper panel shows a schematic view of the genome vector created by defining a bin size of 50bp. In addition, CpGs are shown in a schematic way. A coupling factor is calculated for the centered genomic bin at position *b* (marked by a red vertical line). For this, all CpGs within a maximal distance *d* are considered. The maximal distance d reflects the estimated average size of sequenced DNA fragments. There are several ways for calculating coupling factors. The simplest way is to count the number of CpGs in the surrounding of *b* but with a maximal distance of *d*. Alternatively, a weighting function can be applied to weight each CpG by its distance (*dist*) to the current genomic bin at position *b*. Again, there are several possible weighting functions. The five images at the bottom of the Figure show the progression of the weighting functions *linear*, *exp*, *log*, *count*, and *custom (Down et al. 2008)* by defining *d* = 700.

Whereas the weighting functions *count*, *linear*, *exp*, and *log* are calculated by defined formulas, the custom function allows for specifying user-defined weights for any possible distance *dist*. For example, Down et al. (Down et al. 2008) have generated custom weights for the distances $dist \in [0,648]$. These weights were estimated empirically by sampling from the fragment-length distribution and randomly placing each fragment such that it overlaps the genomic bin Down et al. 2008). Such weights can be up-loaded using MEDIPS and are returned when the *custom* function is called. Let C_{cb} be the coupling factor between a CpG at position *c* and a genomic bin at position *b* calculated based on an arbitrary weighting function and for any specified parameter *d*. Then $C_{tot} = \sum_{c} C_{cb}$ is the sum of coupling factors at the genomic bin *b* with respect to all CpGs at a genomic position *c*, where $|b-c| \leq d$. For simplification, in the following, C_{tot} is called the coupling factor at a genomic bin *b* and gives a measure of local CpG density.

It has been shown Weber et al. 2005; Eckhardt et al. 2006) that in mammalian cells, methylation is negatively correlated to CpG densities. In other words, regions of low CpG density tend to be high methylated, whereas regions of high CpG density tend to be mainly unmethylated. In order to test the correlation of meassured methylation values (Eckhardt et al. 2006) compared to local

CpG densities calculated with respect to the different weighting functions, we have systematically calculated coupling vectors (bin size=50) with varying $d \in [0,2000]$ using the weighting functions *count*, *linear*, *exp*, *log*, as well as for the empirically derived weights presented by Down et al. (Down et al. 2008) (*custom*). Because the custom weights are available for the range $d \in [0,648]$, only, the weight at d=648 is also utilized for the remaining distances up to d=2000. For the comparisons, we have accessed DNA-methylation values derived from bisulphite sequencing experiments of a sperm sample as presented by the human epigenome project (HEP) (Eckhardt et al. 2006). Bisulphite sequencing derived methylation data was generated for approximately 3000 selected genomic regions (called HEP traces) of length 50bp to 500bp (Eckhardt et al. 2006). In order to compare CpG densities to the available methylation data, for all utilized weighting functions with varying parameter d, we have calculated mean coupling factors for each of the HEP traces and examined the relation to corresponding mean methylation values by Pearson correlation. Figure 2a shows the resulting Pearson correlations for varying parameter d and for the several tested weighting functions.

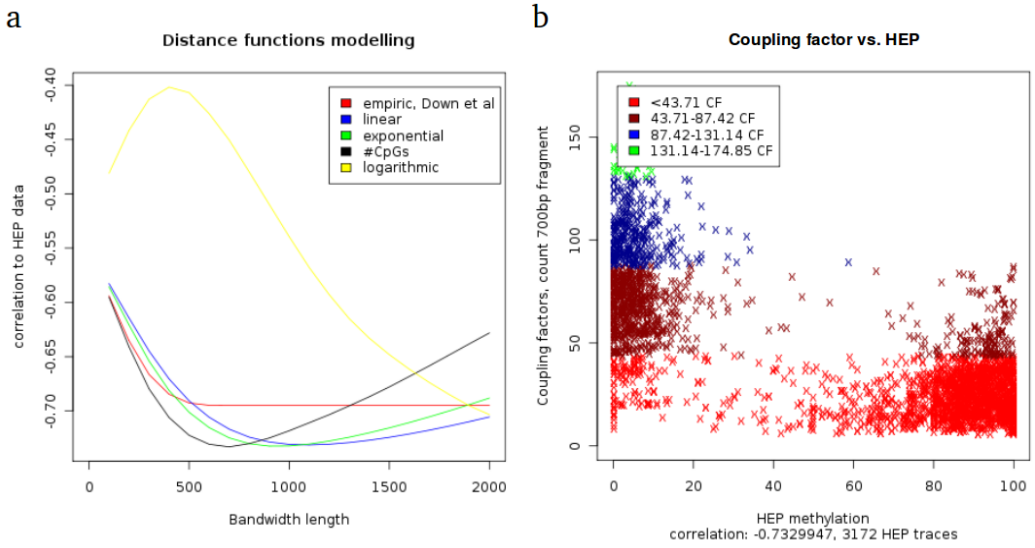

Figure 2: Evaluation of coupling factor calculations. Figure a shows the resulting Pearson correlations (y-axis) between the mean coupling factors and bisulphite sequencing derived mean methylation values for a varying distance parameter *d* (x-axis) and for different weighting factors (colours). The best negative correlation (-0.73) was achieved by setting the parameter *d* = 700 and by using the *count* function. Figure b shows the according scatterplot where each data point represents a HEP trace. The scatterplot contrasts the mean methylation value (x-axis) and mean CpG denisty (y-axis). The color code divides the full range of CpG densities into quantiles.

Interestingly, the best negative correlation (that is the higher the CpG density, the lower the bisulphite derived methylation values) was achieved (-0.73) by setting the parameter $d = 700$ and by using the *count* function. For this parameter settings, Figure 2b shows a scatterplot comparing mean HEP methylation values and mean coupling factors. Here, each data point represents a HEP trace and the plot contrasts the mean methylation value (x-axis) with the mean CpG denisty (y-axis). The color code divides the full range of CpG densities into quantiles. Based on these results, in the following, the coupling vector is calculated by specifying $d = 700$ and by using the *count* function. However, the MEDIPS package allows for justifying the according parameters or for supplying any custom defined distance weights. Moreover, coupling vectors can be calculated for any arbitrary DNA sequence pattern and the resulting coupling vectors can be exported into a WIG file for visualizing the sequence pattern densities along the chromosomes using a suitable genome browser.

2.4.2 Calibration curve

As we have created a genome vector that contains the raw signals at each genomic bin as well as an according coupling vector containing the calculated coupling factors at each genomic bin, the dependency of local MeDIP-seq signal intensities and local CpG densities can be examined. However, by simply plotting the genome vector against the coupling vector, no concrete dependency is observable. However, a dependency between CpG densities and MeDIP-seq signals can be made tangible by calculating the calibration curve (Down et al. 2008). Calculation of the calibration curve is achieved by first dividing the total range of coupling factors into regular levels. Second, all genomic bins are partitioned into these levels by considering their associated coupling factors. Finally, for each level of coupling factors, the mean signal and mean coupling factor of all genomic bins that fall into this level are calculated. As the calibration curve represents the averaged signals and coupling factors over the full range of coupling factors, it reveals the experiment specific dependency between signal intensities and CpG densities (see Supplementary Figures 3a and b of the main manuscript).

In fact, for the low range of coupling factors, the calibration curve indicates that the MeDIP-seq signals, in average, increase because of an increasing CpG density. Therefore, an increased signal is not necessarily caused by a higher level of mCpGs but scales with the general CpG density. In contrast, for INPUT derived sequencing data this dependency of CpG density and sequencing

signals is not observable (see Supplementary Figure 3c of the main manuscript). Therefore, the calibration plot is very characteristic for MeDIP-seq data and the quality of the enrichment step of the MeDIP experiment can be estimated by visual inspection of the progression of the calibration curve. For higher levels of CpG densities, the mean MeDIP-seq signals decrease. It is assumed that this decrease is caused by the fact that in biological systems, regions of higher CpG denstities are mainly unmethylated. Interestingly, in biological systems, cytosine methylation occurs mainly in regions of low CpG density. The other way round, cytosines located in regions of high CpG density are mainly unmethylated. This circumstance implicates that the dependency between increased signal intensities caused by increased CpG densities is visible for regions of low CpG densities, only.

2.4.3 Relative and absolute methylation scores

The calibration curve reveals that, in average, an increase of MeDIP-seq signals is caused by an increasing CpG density. This approximately linear dependency is visible for the low range of coupling factors, only. For higher levels of CpG densities, the mean MeDIP-seq signals decrease. As mentioned above, it is assumed that this decrease is caused by the fact that in mammalian cells, regions of higher CpG denstities are mainly unmethylated. In agreement with this assumption, Pelizzola and colleagues (Pelizzola et al. 2008) have shown that the dependency of MeDIP derived signals and CpG density continues for higher levels of CpG densities, by analysing artificially fully methylated samples using MeDIP-Chip. In fact, they identified a sigmoidal dependency between CpG density and MeDIP-Chip data (Pelizzola et al. 2008). In agreement with Pelizzola et al. (Pelizzola et al. 2008), the signal plateau in the lower range of chip signals is caused by background noise and it is assumed that the signal plateau in the upper range of chip signals occurs by a saturation of hybridization events and is therefore an array specific artefact.

By visual inspection of the MeDIP-seq derived calibration curves, and motivated by the observations made by Pelizzola et al. (Pelizzola et al. 2008), a continuing linear dependency of MeDIP-seq signals for higher levels of CpG densities is assumed. Analogous to Down et al. (Down et al. 2008), the local maximum of mean MeDIP-seq signals of the calibration curve in the lower part of coupling factors is identified. Let

$$y = y_1,\ldots,y_l$$

be the mean coupling factors, and let

$$x = x_1,...,x_l$$

be the according mean MeDIP-seq signals of the calibration curve, where l is the number of tested coupling factor levels and $i = 1,...,l$, then the smallest level i is identified, where

$$x_{i-3}, x_{i-2}, x_{i-1} \leq x_i \geq x_{i+1}, x_{i+2}, x_{i+3}.$$

Let $i_{\max}$ be the according identified level of i, then

$$y_{\max} = y_1,...,y_{i_{\max}}$$

$$x_{\max} = x_1,...,x_{i_{\max}}$$

is the part of the calibration curve in the low range of coupling factors, where an approximately linear dependency between MeDIP-seq signals and coupling factors is observed. Here, $x_{\max}$ can be explained by a function of $y_{\max}$ as

$$x_{\max} = f(y_{\max}) + \varepsilon$$

where ε is an error variable (i.e. measurement errors) that is expected to spread by chance and therefore, its expectation value is $E(\varepsilon) = 0$. Because a linear dependency between $x_{\max}$ and $y_{\max}$ is assumed, $x_{\max}$ can be described as

$$x_{\max} = \alpha + \beta \cdot y_{\max} + \varepsilon$$

where the parameter α is the theoretical y-intercept, and the parameter β is the theoretical slope. Based on the pre-calculated $x_{\max}$ and $y_{\max}$ vectors, linear regression is performed, in order to identify a suitable linear model. Linear regression estimates concrete values a and b for the parameters α and β so that it is valid:

$$x_{\max_i} = a + b \cdot y_{\max_i} + e_i$$

where $i = 1,...,i_{\max}$. Here, the residuum e_i reflects the difference between the regression curve $a + b \cdot y_{\max_i}$ and the measurements for $x_{\max_i}$. Moreover, $x_{\max_i}$ can be replaced by an estimate $\hat{x}_{\max_i}$, where $x_{\max_i} - \hat{x}_{\max_i} = e_i$ and therefore, it is valid:

$$\hat{x}_{\max_i} = a + b \cdot y_{\max_i}$$

MEDIPS calculates the linear regression model using the least squares approach (www.R-Project.org) and concrete values a and b are obtained. Subsequently, for the low range of coupling factors, the observed progression of the calibration curve can be modelled. As discussed

above, a continuing linear dependency between MeDIP-seq signals and CpG density is expected for the higher range of coupling factors. Based on the obtained linear model parameters, concrete $\hat{x}_{\max_i}$ values can be calculated for the full range of coupling factors. Therefore,

$$\hat{x} = \hat{x}_1,...,\hat{x}_{\max_i},...,\hat{x}_l$$

are the estimated mean MeDIP-seq signals over the full range of coupling factor levels *l*.

For MeDIP-seq data normalization, $\hat{x}$ is utilized in order to weight the observed MeDIP-seq signals of the genomic bins with respect to their associated coupling factors. Let (x_{bin_i}, y_{bin_i}) be the raw MeDIP-seq signal of the genomic bin *i* (i.e. the number of overlapping extended short reads), and the pre-calculated coupling factor at the genomic bin *i*, where $i = 1,...,m$ and *m* is the total number of genomic bins, then the normalized relative methylation score is defined as

$$rms_{bin_i} = \log 2(\frac{x_{bin_i} \cdot 10^6}{(a + b \cdot y_{bin_i}) \cdot n}) = \log 2(\frac{x_{bin_i} \cdot 10^6}{\hat{x}_{bin_i} \cdot n})$$

where $\hat{x}_{bin_i} = a + b \cdot y_{bin_i}$ is the estimated weighting parameter obtained by considering the coupling factor y_{bin_i} of the genomic bin *i*, and *n* is the total number of short reads considered for the generation of the genome vector. Based on the total number of short reads (*n*), the raw MeDIP-seq signals are, in parallel, transformed into a reads per million (*rpm*) format in order to assure that *rms* values are comparable between methylomes generated from differing amounts of short reads. The MEDIPS package subsequently transforms the resulting *rms* data range into the consistent interval [0,1000], before finally returned. We consider the *rms* values as the normalized MeDIP-seq signals corrected for the effect of unspecific antibody binding.

In order to identify an absolute methylation estimate for any specified region of interest, i.e. either any functional genomic regions like promoters or CpG islands or genome wide windows of arbitrary length, the raw MeDIP-seq values are normalized into absolute methylation scores (*ams*). The absolute methylation scores correct for the relative CpG density of the regions of interest and therefore, allow for comparing methylation profiles of regions with differing CpG densities. Let $ROI = ((x_{bin_1}, y_{bin_1}),...,(x_{bin_s}, y_{bin_s}))$ be the raw MeDIP-seq signals and coupling factors of adjacent genomic bins *i* that define a region of interest (*ROI*), where $i = 1,...,s$ and *s* is the total number of genomic bins comprised by the *ROI*, then the absolute methylation score for the *ROI* is defined as

$$ams_{ROI} = \log 2\left(\frac{\frac{1}{s}\sum_{i=1}^{s}\frac{x_{bin_i}\cdot 10^6}{(a+b\cdot y_{bin_i})\cdot n}}{\frac{1}{s}\sum_{i=1}^{s} y_{bin_i}}\right)$$

Again, the MEDIPS package subsequently transforms the resulting *ams* data range into the consistent interval [0,1000], before finally returned. Analogous to Pelizzola et al. (Pelizzola et al. 2008), we interpret the *ams* values (Pelizzola et al. (Pelizzola et al. 2008) call them *rms*), as the measure of the normalized methylation that is independent of the CpG density of the corresponding genomic region.

3 Identification of differentially methylated regions (DMRs)

Identification of DMRs is essential for determining local differences in the methylation profiles of diverse biological samples. While there exist several methods for determining statistically significant enriched genomic regions from ChIP-on-Chip (Li et al. 2005; Johnson et al. 2006; Toedling et al. 2007; Chavez et al. 2009) and ChIP-Seq experiments (Boyle et al. 2008; Ji et al. 2008; Valouev et al. 2008; Lun et al. 2009; Rozowsky et al. 2009), the identification of differentially methylated regions from MeDIP-seq data remains insufficiently explored. The main difference between the ChIP-Seq and MeDIP-seq approaches is that TFBSs are of short length (8-16bp) and therefore, ChIP-Seq specific methods intend to identify isolated short genomic regions of high short read enrichments. In contrast, CpGs are spread more widely along the chromosomes and are partly accumulated in CpG islands of length >300bp. Moreover, methylation alterations may occur at few CpG locations, only, and therefore, no sharp TFBSs like ChIP-Seq peaks are expected. Subsequently, in order to identify DMRs, comparatively longer genomic stretches have to be considered and methylation alterations have to be determined in a more sensitive way.

For the identification of DMRs, we propose two alternative approaches. Firstly, it is of interest to specify pre-defined genomic regions of interest (ROIs) like CpG islands, promoters etc., and to specifically compare methylation patterns for these regions. Secondly, it is of interest to calculate differential methylation for genome wide frames of arbitrary length. However, in both cases we call any predefined genomic region as ROI. Here, we present a statistical approach for calculating differential methylation for any predefined ROI, based on sequencing data from two different MeDIP treated samples (Control and Treatment) with respect to an additional input sequencing

data set (Input). Let *C*, *T*, and *I* be the genome vectors generated based on the sequencing data from Control, Treatment, and Input using an arbitrary bin size *b* and let *ROI* be a set of predefined ROIs:

$$ROI = ROI_1,...,ROI_i,...,ROI_n$$

where *n* is the number of ROIs to be tested and the ROI_i's are of length $m_1,...,m_n$. In the following, the identification of DMRs is only supported for any ROI_i of length $m_i \geq 5 \cdot b$. Therefore, each ROI_i consists out of a set of at least five genomic bins (bin_{ROI_i}), where $bin_{ROI_i} = bin_{i,1},...,bin_{i,j},...,bin_{i,k_i} \in ROI_i$ and $k_i = floor(\frac{m_i}{b})$. For each ROI_i, mean *rpm* and *rms* values are calculated based on *C* and *T* as:

$$C.RPM_{ROI_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} rpm(C.bin_{i,j})$$

$$C.RMS_{ROI_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} rms(C.bin_{i,j})$$

$$T.RPM_{ROI_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} rpm(T.bin_{i,j})$$

$$T.RMS_{ROI_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} rms(T.bin_{i,j})$$

where $rpm(C.bin_{i,j})$, $rms(C.bin_{i,j})$, $rpm(T.bin_{i,j})$, and $rms(T.bin_{i,j})$ are the pre-calculated *rpm* (see section 2.2) and *rms* values (see section 2.4.3) for the according genomic bins of the Control and of the Treatment samples. In addition, for each ROI_i, mean *rpm* values are calculated based on *I* as:

$$I.RPM_{ROI_i} = \frac{1}{k_i}\sum_{j=1}^{k_i} rpm(I.bin_{i,j})$$

where $rpm(I.bin_{i,j})$ are the pre-calculated *rpm* values for the genomic bins of the Input sample.

Based on the mean *rms* values of the Control and of the Treatment sample, for each ROI_i the following ratio is calculated:

$$r.rms_{ROI_i} = \frac{C.RMS_{ROI_i}}{T.RMS_{ROI_i}}$$

In addition, by considering the mean *rpm* values of the Control or of the Treatment sample, respectively, the following ratios are calculated with respect to rpm values of the Input sample:

$$r.rpm.C_{ROI_i} = \frac{C.RPM_{ROI_i}}{I.RPM_{ROI_i}}$$

$$r.rpm.T_{ROI_i} = \frac{T.RPM_{ROI_i}}{I.RPM_{ROI_i}}$$

Because local background sequencing signals are variable along the chromosomes due to differing DNA availability, a global background rpm signal threshold is estimated based on the distribution of all calculated $I.RPM_{ROI_i}$ values. This is done by defining a targeted quantile *qt* (e.g. *qt* = 0.95) and by identifying the $I.RPM_{ROI_i}$ value (*t*), where *qt*% of all $I.RPM_{ROI_i}$ values are < *t*. Figure 3 illustrates the distributions of the $I.RPM_{ROI_i}$, $C.RPM_{ROI_i}$, and $T.RPM_{ROI_i}$ values as obtained from the Input, hESCs (Control) and DE (Treatment) samples, when defining regions of interest as overlapping genome wide 500 bp windows, where neighbouring windows overlap by 250 bp. By setting the *qt* parameter to *qt=0.90*, here, an *rpm* threshold *t=0.2566* is obtained from the input $I.RPM_{ROI_i}$ distribution.

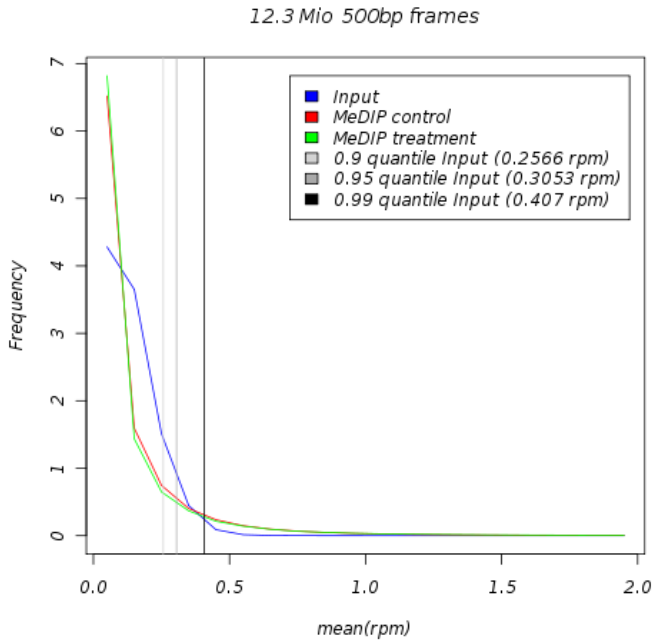

Figure 3: Global mean *rpm* signal distributions. The figure illustrates histograms for the mean *rpm* values of all genome-wide overlapping 500bp windows for hESCs, DE, and input samples. The grey lines indicate three possible global *rpm* thresholds as derived by setting the *qt* parameter to *qt=0.9*, *qt=0.95*, and *qt=0.99*.

This estimated global minimal mean *rpm* threshold *t* will serve as an additional parameter for selecting genomic regions that show mean MeDIP-seq derived *rpm* signals of at least *t* in either the Control or the Treatment sample.

Moreover, statistical testing is utilized in order to rate whether the obtained *rms* data series of the genomic bins within any ROI_i significantly differs in the Control sample compared to the Treatment sample. For each ROI_i it is tested, whether the *rms* values of the genomic bins $bin_{ROI_i} = bin_{i,1},...,bin_{i,j},...,bin_{i,k_i} \in ROI_i$ of the Control sample significantly differ from the *rms* values of the according genomic bins of the Treatment sample. For this, the MEDIPS package utilises the *t.test()* and *wilcox.test()* functions of the R environment (www.R-project.org) with default parameter settings (two-sided tests in both cases). Therefore, for each tested ROI_i two p-values ($ROI.p.value.t_i$ and $ROI.p.value.w_i$) will be calculated and serve as a further level for discriminating between local methylation profiles.

For identifying ROI_i's that show differential methylation between the Control and the Treatment sample and with respect to the Input sample, based on the pre-calculated parameters, a filtering procedure is performed. The following filtering procedure also discriminates between increased

methylation in the Control sample compared to the Treatment sample (Control>Treatment, a) and vice versa (Treatment>Control, b):

1. ROI_i's where $C.RMS_{ROI_i} = T.RMS_{ROI_i} = 0$ are neglected,
2. ROI_i's where $ROI.p.value.t_i > p$ and $ROI.p.value.w_i > p$ are neglected, where *p* is any targeted level of significance,
3. filtering for the ratio:
    a. ROI_i's where $r.rms_{ROI_i} < h$ are neglected, where *h* is an upper ratio threshold,
    b. ROI_i's where $r.rms_{ROI_i} > l$ are neglected, where *l* is a lower ratio threshold,
4. filtering for global Input derived background signals:
    a. ROI_i's where $C.RPM_{ROI_i} < t$ are neglected,
    b. ROI_i's where $T.RPM_{ROI_i} < t$ are neglected,
5. filtering for local Input derived background signals:
    a. ROI_i's where $r.rpm.C_{ROI_i} < h$ are neglected,
    b. ROI_i's where $r.rpm.T_{ROI_i} < h$ are neglected.

The remaining ROI_i are considered as candidate genomic regions where events of differential methylation can be deduced from the data in a sophisticated statistical way.

For selecting significant regions that show de- or *de-novo* methylation events, we executed the *MEDIPS.selectSignificants()* function of the MEDIPS package two times separately, and specified the following parameters: *qt*=0.9, *up*=1.333333; *down*=0.75, *p.value*=0.001. Afterwards, we ended up with highly significant candidate regions of differential methylation. Because we have executed the according *MEDIPS.diffMethyl()* function for overlapping 500bp windows, we partly received overlapping significant frames. Therefore, we finally merged overlapping regions into one super sized region using the *MEDIPS.mergeFrames()* function of the MEDIPS package.

References

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Chavez, L., Bais, A.S., Vingron, M., Lehrach, H., Adjaye, J., and Herwig, R. 2009. In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics* **10**: 314.

Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M. et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**(7): 779-785.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A. et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**(12): 1378-1385.

Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196**(2): 261-282.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10): R80.

Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., and Wong, W.H. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**(11): 1293-1300.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., and Liu, X.S. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A* **103**(33): 12457-12462.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pheasant, M. et al. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**(Database issue): D755-761.

Li, W., Meyer, C.A., and Liu, X.S. 2005. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21 Suppl 1**: i274-282.

Lun, D.S., Sherrid, A., Weiner, B., Sherman, D.R., and Galagan, J.E. 2009. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol* **10**(12): R142.

Pages, H. BSgenome: Infrastructure for Biostrings-based genome data packages.

Pelizzola, M., Koga, Y., Urban, A.E., Krauthammer, M., Weissman, S., Halaban, R., and Molinaro, A.M. 2008. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res* **18**(10): 1652-1659.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M.B. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**(1): 66-75.

Toedling, J., Skylar, O., Krueger, T., Fischer, J.J., Sperling, S., and Huber, W. 2007. Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8**: 221.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**(9): 829-834.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**(8): 853-862.