Bioinformatics Advance Access published September 17, 2012

MeDIP-HMM: Genome-wide identification of distinct DNA methylation states from high-density tiling arrays

Michael Seifert [1,2,3]*, Sandra Cortijo [2], Maria Colomé-Tatché [4], Frank Johannes [4], François Roudier [2] and Vincent Colot [2]

[1]Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
[2]Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique (CNRS) UMR8197, Paris, France
[3]Cellular Networks and Systems Biology, Biotechnology Center of the Technical University Dresden, Dresden, Germany
[4]Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Methylation of cytosines in DNA is an important epigenetic mechanism involved in transcriptional regulation and preservation of genome integrity in a wide range of eukaryotes. Immunoprecipitation of methylated DNA followed by hybridization to genomic tiling arrays (MeDIP-chip) is a cost-effective and sensitive method for methylome analyses. However, existing bioinformatics methods only enable a binary classification into unmethylated and methylated genomic regions, which limits biological interpretations. Indeed, DNA methylation levels can vary substantially within a given DNA fragment depending on the number and degree of methylated cytosines. Therefore, a method for the identification of more than two methylation states is highly desirable.

Results: Here, we present a three-state Hidden Markov Model (MeDIP-HMM) for analyzing MeDIP-chip data. MeDIP-HMM utilizes a higher-order state-transition process improving modeling of spatial dependencies between chromosomal regions, allows a simultaneous analysis of replicates, and enables a differentiation between unmethylated, methylated and highly methylated genomic regions. We train MeDIP-HMM using a Bayesian Baum-Welch algorithm integrating prior knowledge on methylation levels. We apply MeDIP-HMM to the analysis of the Arabidopsis root methylome and systematically investigate the benefit of using higher-order HMMs. Moreover, we also perform an in-depth comparison study to existing methods and demonstrate the value of using MeDIP-HMM by comparisons to current knowledge on the Arabidopsis methylome. We find that MeDIP-HMM is a fast and precise method for the analysis of methylome data enabling the identification of distinct DNA methylation levels. Finally, we provide evidence for the general applicability of MeDIP-HMM by analyzing promoter DNA methylation data obtained for chicken.

Availability: MeDIP-HMM is available as part of the open source Java library Jstacs (www.jstacs.de/index.php/MeDIP-HMM).

Contact: seifert@ipk-gatersleben.de

Supplementary information: Supporting appendices, figures and tables are available from the journal's web site. Data files are available from the Jstacs web site.

*to whom correspondence should be addressed

1 INTRODUCTION

Methylation of genomic DNA is one of the best characterized epigenetic modifications catalyzed by DNA methyltransferases that methylate cytosines at their carbon-5 position (Beck and Rakyan (2008)). In mammals, DNA methylation is found exclusively in the CpG dinucleotide context, except in embryonic stem cells, where a small proportion of cytosines in other contexts (CpA, CpT and CpC) are also methylated. In plants, DNA methylation is found in symmetric CpG and CpHpG and in asymmetric CpHpH contexts (H = A, T or C). Generally, DNA methylation plays important roles in regulation of gene expression (Esteller (2007); Zilberman *et al.* (2007); Wutz (2011); Barlow (2011)) and silencing of transposons (Law and Jacobsen (2010); Teixeira and Colot (2010)).

Despite recent developments of next-generation sequencing approaches for determining methylomes at single base pair resolution (e.g. Cokus *et al.* (2008); Lister *et al.* (2008, 2009)), methylomes of different organisms or cell tissues are also frequently analyzed using whole-genome tiling arrays (e.g. Zilberman *et al.* (2008); Borgel *et al.* (2010); Nätt *et al.* (2012)), which provide cost-effective alternatives. Most array-based studies are done based on methylated DNA immunoprecipitation coupled with hybridization to a tiling array (MeDIP-chip) (Beck and Rakyan (2008); Harrison and Parle-McDermott (2011)). MeDIP-chip enables to analyze the methylome of a genome at a resolution of few hundred base pairs, which in most applications is sufficient to draw biologically meaningful conclusions.

The analysis of MeDIP-chip data puts similar challenges on bioinformatics methods as identified for the analysis of closely related array-based chromatin immunoprecipitation data (ChIP-chip). Different methods for the analysis of ChIP-chip data were proposed over the last years. Especially methods based on Hidden Markov Models (HMMs) (e.g Ji and Wong (2005); Humburg *et al.* (2008); Seifert *et al.* (2009)) and methods based on mixture models (e.g. Martin-Magniette *et al.* (2008); Johannes *et al.* (2010); Banaei *et al.* (2011)) were shown to enable reliable predictions of chromosomal target regions of transcription factors or histone modifications. A common characteristic of all these methods is the modeling of two different populations of measurements to differentiate non-enriched genomic regions from enriched

ones. Main conceptual differences exist in the way of modeling dependencies between adjacent measurements on a chromosome, in handling of replicates and in training algorithms. From these methods only approaches based on HMMs integrate dependencies between directly adjacent measurements on a chromosome. Some methods only enable a separate analysis of replicates (e.g. Humburg *et al.* (2008); Martin-Magniette *et al.* (2008); Seifert *et al.* (2009)), whereas others try to improve the analysis by simultaneous modeling of replicates (e.g. Ji and Wong (2005); Johannes *et al.* (2010); Banaei *et al.* (2011)). Mixture models are typically trained by specifically designed Expectation Maximization (EM) algorithms (Dempster *et al.* (1977)). HMM-based methods except TileMAPHMM (Ji and Wong (2005)) and HMMs by Seifert *et al.* (2009) use a Baum-Welch training (Baum (1972)) representing a special case of an EM algorithm. TileMAPHMM is based on data-dependent ad hoc settings, and HMMs by Seifert *et al.* (2009) enable the integration of prior knowledge on measurements using a Bayesian Baum-Welch algorithm.

All these methods are useful tools for the analysis of ChIP-chip data. Additionally, some methods like ChIPmix (Martin-Magniette *et al.* (2008)) and a mixture model approach by Johannes *et al.* (2010) have already been applied to the analysis of MeDIP-chip data. One general limitation of all these methods in the context of MeDIP-chip data analyses is that they only enable a binary classification into unmethylated and methylated regions. MeDIP-chip data is known to be more complex showing differences in methylation levels of individual chromosomal regions as for example revealed for *Arabidopsis thaliana* having moderately methylated genes and highly methylated transposons (Zilberman *et al.* (2008)). To address that, a three-state HMM specifically designed for the analysis of Arabidopsis MeDIP-chip data has been developed in a companion work by Cortijo et al. (2012). This approach utilizes Arabidopsis-specific ad hoc settings and enabled a better interpretation of MeDIP-chip data using a classification of methylation states of genomic regions according to the underlying three-states. Generally, organism-specific ad hoc settings can only be hardly transferred to MeDIP-chip data of other organisms. Thus, there is still a great demand of having a general method that is able to differentiate between different methylation levels without being dependent on a specific organism. Moreover, a systematic performance evaluation of different methods for analyzing DNA methylation data has not been carried out up to now, and current HMM-based methods only focus on standard first-order HMMs.

Here we present MeDIP-HMM, a method specifically designed for the analysis of MeDIP-chip data. MeDIP-HMM utilizes three states to differentiate between unmethylated, methylated and highly methylated regions overcoming limitations of typically used ChIP-chip methods only enabling a binary classification into unmethylated and methylated regions. Additionally, MeDIP-HMM can perform a simultaneous analysis of replicates and integrates prior knowledge on measurements to improve the identification of methylated genomic regions. Moreover, MeDIP-HMM can also take advantage of higher-order hidden Markov chains to improve spatial modeling of dependencies between neighboring regions. This has recently been found to improve the analysis of comparative genomics data (Seifert *et al.* (2012)) and provides a valuable option for improving the analysis of MeDIP-chip data. We apply our MeDIP-HMM to the analysis of the Arabidopsis root methylome and systematically evaluate the influence of using higher-order Markov chains on the identification of methylated genomic regions. We further perform an in-depth comparison study to widely used existing methods and demonstrate advantages of using MeDIP-HMM based on comparisons to current knowledge on the Arabidopsis methylome. We also show that MeDIP-HMM can be applied to non-Arabidopsis data by performing an additional study on promoter DNA methylation data obtained for chicken (Nätt *et al.* (2012)).

2 METHODS

In this section, we initially describe our root methylome data. Then we provide the mathematical background of MeDIP-HMM. Finally, we consider publicly available data for model evaluations.

2.1 Arabidopsis root methylome data set

We performed a MeDIP-chip experiment to identify genomic regions that are methylated in root tissue of the accession Col-0 of the flowering plant *A. thaliana* according to the experimental protocol described by Cortijo *et al.* (2012). The data set is publicly available from GEO (GSE36750). This data set represents log-ratios of fluorescent intensities of immunoprecipitated methylated DNA versus reference input DNA measured for $T := 711,320$ genomic regions in two biological replicates. We applied quantile normalization (Bolstad *et al.* (2003)) to the log-ratios of both replicates and summarized the resulting normalized log-ratios of each replicate in chromosome-specific methylation profiles. This leads to a methylation profile $\vec{o}(k) := (\vec{o}_1(k), \ldots, \vec{o}_{T_k}(k))$ for each chromosome $k \in \{1, \ldots, 5\}$ containing chromosome-specific measurements $\vec{o}_t(k) := (o_t^1(k), o_t^2(k))$ of both replicates in increasing order of their chromosomal positions. Thus, each methylation level $\vec{o}_t(k)$ of a region $t \in \{1, \ldots, T_k\}$ is represented by the corresponding normalized log-ratios $o_t^1(k)$ and $o_t^2(k)$ measured in replicates 1 and 2, respectively. An histogram of average methylation levels is shown in Figure 1a. Measurements of both biological replicates are highly reproducible reaching a Pearson correlation coefficient of 0.92 (Figure S1a).

2.2 MeDIP-HMM: Hidden Markov Model for MeDIP-chip analyses

We utilize a three-state HMM with state-specific multivariate Gaussian emission densities to analyze methylation levels of chromosomal regions in methylation profiles. Motivated by the distribution of methylation levels in Figure 1a, three states $S := \{'\mathrm{U}', '\mathrm{M}', '\mathrm{I}'\}$ are defined to model distinct classes of methylation states. State $'\mathrm{U}'$ models unmethylated regions characterized by log-ratios of about or much less than zero. Highly methylated regions having log-ratios much greater than zero are modeled by state $'\mathrm{M}'$. Methylated regions having log-ratios in between unmethylated and highly methylated genomic regions are modeled by state 'I'. These states are the basis of the fully connected three-state architecture of the HMM shown in Figure S2.

More formally, the state of a region t on chromosome k is denoted by $q_t(k) \in S$. To account for correlations between measurements of closely adjacent regions on a chromosome, a state sequence $\vec{q}(k) := (q_1(k), \ldots, q_{T_k}(k))$ underlying a methylation profile $\vec{o}(k)$ is modeled by a homogeneous Markov model of order L (e.g. Berchtold and Raftery (2002)). Thus, the state-transition process of an HMM of order $L \geq 1$ is parameterized by an

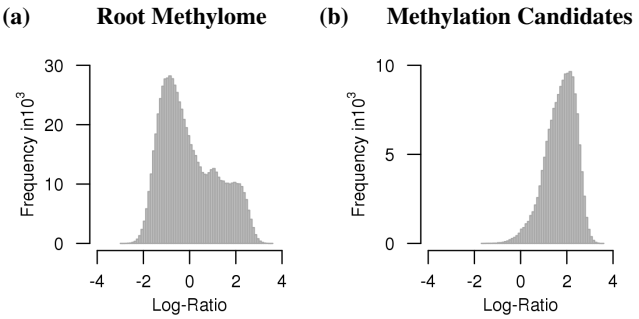

Fig. 1. Overview of measurements in the Arabidopsis root methylome data set. **(a)** Histogram of average methylation levels measured in root tissue using two biological replicates. Three groups of methylation levels are observed comprising unmethylated genomic regions with log-ratios much less than zero peaking around -1.0, highly methylated genomic regions with log-ratios much greater than zero peaking around 2.0, and genomic regions having methylation levels between unmethylated and highly methylated regions peaking around 1.0. **(b)** Histogram of average methylation levels of genomic regions in the root methylome data set that have been labeled as potential candidates for DNA methylation based on an initial study by Zilberman *et al.* (2008). The distribution of measured methylation levels for the candidate regions is clearly shifted into the positive range of log-ratios peaking at log-ratios of about 2.0. This strongly indicates that candidate regions of DNA methylation from Zilberman *et al.* (2008) are also present in our root methylome data set. This motivates the usage of these information for evaluating different methods for MeDIP-chip analyses.

initial state distribution $\vec{\pi} := (\pi_i)_{i \in S}$ with initial state probability $\pi_i \in (0,1)$ and a set of stochastic transition matrices $A := \{A_1, \ldots, A_L\}$. The initial state distribution fulfills the constraint $\sum_{i \in S} \pi_i = 1$. Each transition matrix $A_l := (a_{ij})_{i \in S^l, j \in S} \in A$ with $1 \leq l \leq L$ specifies the transition probability $a_{ij} \in (0,1)$ for each transition from the current state i_l of a state-context $i = (i_1, \ldots, i_l) \in S^l$ to a next state $j \in S$. Thus, this means that for $l > 1$ transitions from i_l are depending on its $l-1$ predecessors $i_1, \ldots, i_{l-1}$. Hence, a transition matrix A_l with $1 \leq l < L$ is used for the transition from the current state $q_l(k)$ to the next state $q_{l+1}(k)$ under consideration of the $l-1$ predecessor states $q_1(k), \ldots, q_{l-1}(k)$. The transition matrix $A_L \in A$ is used for each transition from $q_t(k)$ to $q_{t+1}(k)$ for all regions $t \geq L$ in dependency of the complete memory on $L-1$ predecessor states $q_{t-L+1}(k), \ldots, q_{t-1}(k)$. Finally, each $A_l \in A$ also fulfills the constraint $\sum_{j \in S} a_{ij} = 1$ for each $i \in S^l$.

Practically, only small model orders should be considered due to an exponential increase of transition parameters with increasing model order leading to higher computational complexities and potentially overfitted models. Thus, different studies in other domains mainly focused on second-order HMMs (e.g. Mari *et al.* (1997); Eng *et al.* (2009)) or developed different strategies to obtain parsimonious models (e.g. du Preez (1998); Wang (2006); Seifert *et al.* (2012)). For our approach, the most parsimonious model is obtained for $L = 0$. This reduces the HMM to a mixture model (e.g. Bilmes (1998)) that does not model dependencies between measurements.

Generally, the state sequence $\vec{q}(k)$ underlying a methylation profile $\vec{o}(k)$ is unknown. To enable the inference of a state sequence, measurements contained in a methylation profile must be integrated into the HMM by making use of state-specific emission distributions. The usage of univariate Gaussian emission distributions represents a common choice for HMM-based analyses of single ChIP-chip experiments (e.g. Li *et al.* (2005); Seifert *et al.* (2009)). Similar to Johannes *et al.* (2010), we extend this assumption by using multivariate Gaussian emission distributions enabling the simultaneous analysis of replicates of an experiment. Thus, a methylation level $\vec{o} := (o^1, \ldots, o^d)$ that represents the log-ratios of a region measured in d replicates is modeled under state $i \in S$ of the HMM by the state-specific Gaussian emission distribution

$$b_i(\vec{o}) := \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp\left(-\frac{1}{2}(\vec{o} - \vec{\mu}_i) \cdot \Sigma_i^{-1} \cdot (\vec{o} - \vec{\mu}_i)^T\right)$$

with mean vector $\vec{\mu}_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$. Here, the determinant of the covariance matrix is denoted by $\det(\Sigma_i)$, Σ_i^{-1} represents the inverse of the covariance matrix, and the transpose vector of $(\vec{o} - \vec{\mu}_i)$ is given by $(\vec{o} - \vec{\mu}_i)^T$. The emission parameters of all states of the HMM are summarized by $B := (\vec{\mu}_i, \Sigma_i)_{i \in S}$. All parameters of the HMM are denoted by $\lambda := (\vec{\pi}, A, B)$. The underlying state space model highlighting dependencies modeled between states and emissions is illustrated in Figure S3 for a second-order HMM.

To quantify the methylation status of a genomic region, the HMM is used to compute the probability that a region t in methylation profile $\vec{o}(k)$ is modeled by a state $i \in S$. The corresponding state-posterior probability $\gamma_t^k(i) := P[\, q_t(k) = i \,|\, \vec{o}(k), \lambda \,]$ is computed by the Forward-Backward algorithm adapted to higher-order HMMs (e.g. Seifert (2010)). These state-posterior probabilities allow to decode the most likely underlying methylation state of each region. Additionally, the state-posterior probabilities enable a ranking of genomic regions according to their probabilities of being methylated ('I' or 'M') using the probability $1 - \gamma_t^k(\text{'U'})$ as score.

2.3 Integration of prior knowledge

The integration of prior knowledge on the distribution of measurements enables a problem-specific characterization of model parameters. Especially, the modeling of prior knowledge on emission parameters can substantially improve HMM-based predictions compared to predictions of HMMs ignoring prior knowledge during training (Seifert *et al.* (2011)). For that reason, a prior distribution for an HMM $\lambda := (\vec{\pi}, A, B)$ is defined by

$$P[\, \lambda \,|\, \Theta \,] := D_1(\, \vec{\pi} \,|\, \Theta_1 \,) \cdot D_2(\, A \,|\, \Theta_2 \,) \cdot D_3(\, B \,|\, \Theta_3 \,) \quad (1)$$

given specific hyperparameters $\Theta := (\Theta_1, \Theta_2, \Theta_3)$. This prior represents a product of independent conjugate priors for each class of model parameters enabling analytical parameter estimations and integration of prior knowledge during model training.

Following the usual choice of prior distributions for initial state and transition parameters (e.g. Durbin *et al.* (1998); Seifert *et al.* (2011, 2012)), the prior $D_1(\, \vec{\pi} \,|\, \Theta_1 \,)$ for the initial state distribution is given by a Dirichlet distribution and the prior $D_2(\, A \,|\, \Theta_2 \,)$ for the set of transition matrices is specified by products of Dirichlet distributions. Appendix A of the supplementary data provides details to both prior distributions and chosen hyperparameters.

For the state-specific multivariate Gaussian emission densities enabling simultaneous modeling of measurements of replicates of an experiment, we utilize a Gaussian-Wishart prior to integrate prior

knowledge on different methylation levels. This choice is motivated by Gauvain and Lee (1994) introducing this prior into HMM-based speech recognition. We transfer this to HMM-based modeling of multivariate MeDIP-chip data. Thus, the prior distribution for the emission parameters is a product of state-specific independent Gaussian-Wishart distributions defined by

$$D_3(\,B\,|\,\Theta_3\,) \propto \prod_{i\in S} \det(\Sigma_i^{-1})^{\frac{r_i-d}{2}} \cdot \exp\left(-\frac{1}{2}\,\mathrm{tr}(\Omega_i \cdot \Sigma_i^{-1})\right) \cdot \exp\left(-\frac{\epsilon_i}{2}(\vec{\mu}_i - \vec{\eta}_i)\cdot\Sigma_i^{-1}\cdot(\vec{\mu}_i-\vec{\eta}_i)^T\right)$$

with hyperparameters $\Theta_3 := (\vec{\eta}_i, \epsilon_i, \Omega_i, r_i)_{i\in S}$. Here, $\vec{\eta}_i \in \mathbb{R}^d$ specifies an a priori mean vector for methylation levels modeled by state $i \in S$, and $\epsilon_i \in \mathbb{R}^+$ defines a corresponding scaling factor of the a priori mean vector weighting its strength of influence on the state-specific mean vector $\vec{\mu}_i$ during training. Similarly, $\Omega_i \in \mathbb{R}^{d\times d}$ is a positive definite scale matrix for the covariance matrix Σ_i of state i, and $r_i > d-1$ is a scaling parameter for Σ_i. Additionally, $\mathrm{tr}(\Omega_i \cdot \Sigma_i^{-1})$ specifies the trace of the matrix product $\Omega_i \cdot \Sigma_i^{-1}$.

The influence of the emission prior on the estimation of emission parameters is shown in the following section. Details to chosen prior hyperparameters for the analysis of the MeDIP-chip data are given in the section model initialization.

2.4 Bayesian Baum-Welch training

The adaptation of the HMM-parameters to the MeDIP-chip data is done by a Bayesian Baum-Welch training outlined in Seifert *et al.* (2011) for HMMs with univariate Gaussian emission distributions. This training algorithm is a special case of an EM algorithm (Dempster *et al.* (1977)) enabling the integration of prior knowledge. Based on a choice of initial model parameters, the Bayesian Baum-Welch algorithm iteratively maximizes the posterior density (product of likelihood and prior distribution) of the model parameters given a data set reaching at least a local optimum in dependency of the initial parameters. This is usually done in log-space by successively computing updated HMM-parameters

$$\lambda(h+1) = \underset{\lambda}{argmax}\ (Q(\,\lambda\,|\,\lambda(h)\,) + \log(P[\,\lambda\,|\,\Theta\,]))$$

maximizing the sum of Baum's auxiliary function $Q(\,\lambda\,|\,\lambda(h)\,)$ (see Appendix B) and the logarithm of the prior distribution $P[\,\lambda\,|\,\Theta\,]$ in Eq. (1). This enables an iterative estimation of the new HMM-parameters $\lambda(h+1)$ under consideration of current parameters $\lambda(h)$ (for $h = 1$ initial HMM) until the log-posterior grows less than a pre-defined threshold. Here, we use a threshold of 10^{-3} for two successive iterations.

Details to the estimation of initial state and transition parameters are given in Appendix B of the supplementary data. Since we make use of state-specific multivariate Gaussian emission densities in combination with a Gaussian-Wishart prior, estimation formulas of the state-specific mean vector $\vec{\mu}_i$ and the corresponding covariance matrix Σ_i are given in the following. Detailed derivations of these formulas are outlined in Appendix B of the supplementary data.

Considering the iteration step h of the Bayesian Baum-Welch training, the mean vector of the multivariate Gaussian emission density of state $i \in S$ is given by

$$\vec{\mu}_i^{(h+1)} := \frac{\left(\sum_{k=1}^{K}\sum_{t=1}^{T_k} \vec{o}_t(k)\cdot\gamma_t^k(i)\right) + \epsilon_i\vec{\eta}_i}{\left(\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_t^k(i)\right) + \epsilon_i}$$

with respect to the measured methylation level $\vec{o}_t(k)$, the state-posterior probability $\gamma_t^k(i) := P[\,q_t(k) = i\,|\,\vec{o}(k), \lambda(h)\,]$ under the current model $\lambda(h)$, and the state-specific a priori mean vector $\vec{\eta}_i$ with its scaling factor ϵ_i specified by the Gaussian-Wishart prior. The corresponding state-specific covariance matrix

$$\Sigma_i^{(h+1)} := \frac{\left(\sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_t^k(i)\cdot O_{itk}\right) + \epsilon_i V_i + \Omega_i}{\left(\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_t^k(i)\right) + r_i - d}$$

is computed based on the state-posterior $\gamma_t^k(i)$ under the current model $\lambda(h)$ and two state-specific matrices $O_{itk} := (\vec{o}_t(k) - \vec{\mu}_i^{(h+1)})^T \cdot (\vec{o}_t(k) - \vec{\mu}_i^{(h+1)}) \in \mathbb{R}^{d\times d}$ and $V_i := (\vec{\mu}_i^{(h+1)} - \vec{\eta}_i)^T \cdot (\vec{\mu}_i^{(h+1)} - \vec{\eta}_i) \in \mathbb{R}^{d\times d}$. The a priori mean vector $\vec{\eta}_i$, the scale matrix Ω_i and the scaling factor r_i are specified by the state-specific Gaussian-Wishart prior.

The obtained parameter estimation formulas for the mean vector and the covariance matrix generalize the estimation formulas for multivariate Gaussian emission densities typically used in the standard Baum-Welch training that does not enable the integration of prior knowledge (e.g. Bilmes (1998)). A computational scheme summarizing the main steps of the Bayesian Baum-Welch training is given in Appendix B of the supplementary data.

2.5 Model initialization

To enable the identification of distinct DNA methylation states from MeDIP-chip data, an initial HMM is specified in a data-dependent manner. We use the following heuristic approach to set the initial model parameters.

Initially, the user has to specify the model order L and rough estimates of expected proportions $\pi_{'\mathrm{U}'} \in (0,1)$ and $\pi_{'\mathrm{M}'} \in (0,1)$ of unmethylated and highly methylated genomic regions in an experiment. Here, an histogram or a cumulative density plot of measured methylation levels helps to select these proportions. Alternatively, these proportions can also be chosen based on prior knowledge from previous experiments. Based on that, the initial state distribution is set to $\vec{\pi} := (\pi_{'\mathrm{U}'}, \pi_{'\mathrm{I}'}, \pi_{'\mathrm{M}'})$ with $\pi_{'\mathrm{I}'} := 1 - \pi_{'\mathrm{U}'} - \pi_{'\mathrm{M}'}$. Additionally, the initial transition matrix $A_1 := (a_{ij})_{i,j\in S}$ is defined to have a stationary distribution identical to $\vec{\pi}$ by using state-specific diagonal and non-diagonal elements $a_{ii} := 1 - s/\pi_i$ and $a_{ij} := s/(2\pi_i)$ with respect to $s \in (0, \min\{\pi_{'\mathrm{U}'}, \pi_{'\mathrm{I}'}, \pi_{'\mathrm{M}'}\})$ for controlling the state durations (default $s = 0.05$). For transition matrices $A_l := (a_{ij})_{i\in S^l, j\in S}$ with $1 < l \leq L$, we initially set $a_{ij} := a_{i_l j}$ for each state-context $i := (i_1, \ldots, i_l)$ to the value of the corresponding transition probability $a_{i_l j}$ defined for A_1. These settings realize that the state-transition process of the initial HMM is modeling the specified proportions of unmethylated and highly methylated genomic regions.

In addition to this, the states of the initial HMM need to be characterized by specific Gaussian emission densities to enable the differentiation of methylation levels. For realizing this, average methylation levels of genomic regions are initially computed based on all replicates of an experiment. The resulting distribution of average methylation levels is further divided into three parts by computing data-dependent quantiles $Q_{\pi_{'\mathrm{U}'}}$ and $Q_{\pi_{'\mathrm{U}'}+\pi_{'\mathrm{I}'}}$ for the corresponding initial proportions $\pi_{'\mathrm{U}'}$ and $\pi_{'\mathrm{I}'}$. These two quantiles are used to obtain the following partitioning in which unmethylated genomic regions are assumed to have average methylation levels less than $Q_{\pi_{'\mathrm{U}'}}$, highly methylated genomic regions are assumed to have average methylation levels greater than $Q_{\pi_{'\mathrm{U}'}+\pi_{'\mathrm{I}'}}$, and less strongly methylated genomic regions are assumed to have average methylation levels between $Q_{\pi_{'\mathrm{U}'}}$ and $Q_{\pi_{'\mathrm{U}'}+\pi_{'\mathrm{I}'}}$. For these three groups, the mean values $\mu_{'\mathrm{U}'}$, $\mu_{'\mathrm{I}'}$ and $\mu_{'\mathrm{M}'}$, and the standard deviations $\sigma_{'\mathrm{U}'}$, $\sigma_{'\mathrm{I}'}$ and $\sigma_{'\mathrm{M}'}$ are computed for the corresponding averaged methylation levels. Additionally, Pearson's correlation coefficients $R_{'\mathrm{U}'}(v,w)$, $R_{'\mathrm{I}'}(v,w)$ and $R_{'\mathrm{M}'}(v,w)$ of methylation levels between each pair of replicates (v,w) with $1 \leq v,w \leq d$ and $v \neq w$ are computed for the three groups. Based on these precomputations, the initial mean vector $\vec{\mu}_i$ of each state $i \in S$ is set to $\vec{\mu}_i := (\mu_i,\ldots,\mu_i)$ using the precomputed mean value μ_i. The corresponding covariance matrix $\Sigma_i := (\sigma_i(v,w))$ with $1 \leq v,w \leq d$ is specified by diagonal elements $\sigma_i(v,v) := \sigma_i^2$ and by non-diagonal elements $\sigma_i(v,w) := \sigma_i^2 \cdot R_i(v,w)$ based on the precomputed standard deviation σ_i and the correlation coefficient $R_i(v,w)$. These initial emission parameters realize an appropriate characterization of the three HMM states for identifying distinct classes of methylation levels.

All initially chosen model parameters are further refined during the Bayesian Baum-Welch training using precomputed data-dependent prior knowledge. This is done by setting each a priori mean vector $\vec{\eta}_i := \vec{\mu}_i$ for modeling the methylation levels under state $i \in S$ to the initially computed state-specific mean vector $\vec{\mu}_i$. The corresponding scaling factor is specified by $\epsilon_i := \pi_i \cdot T$ representing the number of measurements initially assumed to be modeled by state i. The scale matrix of the covariance matrix of state i is set to $\Omega_i := T/100 \cdot \Sigma_i$ in dependency of the precomputed covariance matrix Σ_i weighted by one percent of the total number of measurements. The corresponding scaling parameter is set to $r_i := \pi_i \cdot T$.

The approach for setting the initial parameters of the HMM and for specifying the parameters of the prior distribution has been tested on different MeDIP-chip data sets of root and shoot tissue for varying user-specified proportions $\pi_{'\mathrm{U}'}$ and $\pi_{'\mathrm{M}'}$. The performance of the resulting identification of methylation levels by HMMs trained based on these initial settings was found to be very robust (e.g. Table S1 for four different initializations on root data). Some more general hints considering the initialization are summarized in Appendix C of the supplementary data.

Motivated by the mapping of candidate regions of DNA methylation from Zilberman *et al.* (2008) to our tiling array, we use $\pi_{'\mathrm{U}'} = 0.8$ and $\pi_{'\mathrm{M}'} = 0.1$ for all studies and investigate models of order zero up to four.

2.6 Publicly available data for model evaluations

The methylome of Arabidopsis roots has initially been studied on a genome-wide scale by Zilberman *et al.* (2008). This study used a tiling array with 382,178 tiles, which is less dense than the tiling array platform that we used for our experiments. Still, genomic regions quantified as being methylated in this study provide a useful resource for evaluating the performance of different methods for MeDIP-chip analyses, because they also analyzed the accession Col-0. For that purpose, we downloaded the corresponding two DNA methylation profiles from GEO (GSM307382, GSM307384) and applied quantile normalization to the log-ratios of both experiments. We then averaged both log-ratios measured for each genomic region and determined all chromosomal segments consisting of successive genomic regions with an average log-ratio greater or equal than one. This restrictive log-ratio cutoff ensures that only strongly enriched genomic segments are considered as potential targets of DNA methylation. This led to $18,061$ potential candidate segments of DNA methylation widespread across all chromosomes of *A. thaliana*. These segments were mapped back to the $711,320$ genomic regions present on our tiling array. Based on this mapping, each genomic region that was at least partially covered by one of the candidate segments has been labeled as a putative candidate for DNA methylation in root tissue. This resulted in $156,091$ genomic regions being potential candidates for DNA methylation (22% of regions present on our tiling array). The remaining $555,229$ genomic regions are potential candidates for being unmethylated.

To demonstrate the value of using this public data set to evaluate potential candidates of DNA methylation in our root methylome data, the distribution of measured methylation levels of genomic regions labeled as being methylated is shown in Figure 1b (see Figure S1b for a bivariate density plot). Most of these genomic regions have log-ratios clearly greater than zero peaking at about 2.0. Thus, genomic segments identified as being potential candidates of DNA methylation in the Zilberman *et al.* (2008) data are also present in our root methylome data set and can be utilized as a useful resource for comparisons of different MeDIP-chip analysis methods.

3 RESULTS AND DISCUSSION

In this section, we first investigate the effect of using higher-order MeDIP-HMMs for analyzing the Arabidopsis root methylome. We next perform a systematic comparison study to existing methods and analyze predictions of MeDIP-HMM in the context of the Arabidopsis genome annotation. Finally, we show an application of MeDIP-HMM to publicly available promoter DNA methylation data obtained for chicken.

3.1 Comparison of MeDIP-HMMs of different model orders

To compare the influence of different model orders on the identification of methylated genomic regions by MeDIP-HMMs, we analyzed our root methylome data set with respect to methylated regions from Zilberman *et al.* (2008). We initially trained MeDIP-HMMs of orders zero up to four on our root methylome data. For each model, we next ranked all genomic regions according to their probabilities of being methylated ($'\mathrm{I}'$ or $'\mathrm{M}'$) as described in the methods section. Based on that, we computed for each MeDIP-HMM the corresponding true-positive-rates (TPRs) of identified methylated regions reached at different levels of fixed false-positive-rates (FPRs). The performance of the different MeDIP-HMMs at small FPRs is shown in Figure 2.

The largest improvement in the identification of known candidate regions of DNA methylation is achieved by the transition from order zero to order one. The zeroth-order MeDIP-HMM represents a mixture model of multivariate Gaussians that does not enable the modeling of dependencies between measurements in close chromosomal proximity. This is overcome by the first-order MeDIP-HMM capable of modeling dependencies between measurements of directly adjacent chromosomal regions. Moreover, an additional increase in performance is reached by the second-order MeDIP-HMM that extends the first-order model by realizing dependencies between two directly adjacent regions to identify the state of the next region. This is exemplarily shown in Figure 2a for a fixed FPR of 1.5% and also observed among the small FPRs considered in Figure 2b. MeDIP-HMMs of order three and four did not reach the performance of the second-order model. These two models identified methylated regions only slightly better than the first-order MeDIP-HMM potentially due to overfitting caused by exponentially growing numbers of transition parameters. All these findings are also supported by performance evaluations based on different training and test sets (Table S2).

In summary, the transition from a mixture model to an HMM has led to the largest improvement in the identification of methylated genomic regions. Overall, the second-order MeDIP-HMM reached the best performance among all considered models. We focus on this model in the following studies.

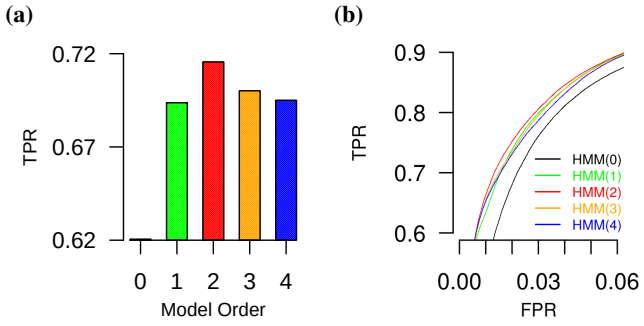

Fig. 2. Evaluation of identified known candidate regions of methylation by MeDIP-HMMs of different model orders. **(a)** True-positive-rates (TPRs) obtained at a fixed false-positive-rate (FPR) of 1.5%. The greatest improvement is reached for the transition from a zero-order model to a first-order model leading to more than 7% increase in TPR. The second-order MeDIP-HMM reaches the best TPR among all models. **(b)** Part of receiver-operating-characteristic (ROC) curves up to a FPR of 6% comparing different MeDIP-HMMs. HMM(L) denotes the corresponding MeDIP-HMM of order L. The second-order MeDIP-HMM (red) reaches the best TPRs at small FPRs.

3.2 Comparison of MeDIP-HMM to existing methods

To compare the second-order MeDIP-HMM against other existing methods, we again utilize our root methylome data set and known potential candidate regions of DNA methylation obtained from Zilberman *et al.* (2008). In recent years, especially methods based on mixture models (e.g. Martin-Magniette *et al.* (2008); Johannes *et al.* (2010); Banaei *et al.* (2011)) and methods based on HMMs (e.g. Ji and Wong (2005); Humburg *et al.* (2008); Seifert *et al.* (2009)) were developed for the analysis of ChIP-chip data to enable reliable identifications of chromosomal target regions of transcription factors or histone modifications. Since the analysis of ChIP-chip data is closely related to that of MeDIP-chip data and because Johannes *et al.* (2010) have already shown that mixture models can be used to analyze DNA methylation data, we compare the identification of known candidate regions of methylation against these different methods. We also include the three-state HMM specifically developed for the analysis of Arabidopsis MeDIP-chip data described in a companion work by Cortijo *et al.* (2012) into the comparison. This model utilizes the observation that introns of protein coding genes are usually unmethylated in Arabidopsis. Based on that, MeDIP-chip data is rescaled and state-specific emission functions are estimated with respect to biological constraints customized for Arabidopsis. Two state-specific Gaussian emission densities with fixed means and equal variances are used to model probes with high or low methylation levels, whereas a mixture of thirty Gaussians with fixed variances is used to model unmethylated probes based on the measurement distribution obtained from intronic probes. The estimation of the thirty Gaussians is done externally and the obtained parameters are used as fixed emission parameters for the unmethylated state. This approach cannot be directly transferred to other organisms like human, where introns are usually found to be methylated (Lister *et al.* (2009)).

The first four columns of Table 1 provide more detailed information about the considered methods. Conceptionally, our MeDIP-HMM is specifically designed for the analysis of MeDIP-chip data by making use of three distinct states to enable the differentiation between different degrees of methylation. Except for CortijoHMM (Cortijo *et al.* (2012)), this cannot be realized by all the other methods that only perform a binary classification into unmethylated and methylated regions. Additionally, MeDIP-HMM also enables the modeling of higher-order dependencies, which has not been addressed by other methods so far.

All methods listed in Table 1 were analyzed for their ability to identify the potential candidate regions of DNA methylation in our root methylome data set. The methods were adapted to the data using their standard initialization and training algorithms. User-defined settings were required to obtain an initial StandardHMM (Seifert *et al.* (2009)) and an initial JohannesModel1 (Johannes *et al.* (2010)). For both models, the initial parameter settings of the first-order MeDIP-HMM have been transferred by taking into account that these models only differentiate between unmethylated and methylated regions. This was done using average initial parameter values of states ′I′ and ′M′ modeling methylations in MeDIP-HMM to initially specify the state representing methylated regions in StandardHMM and JohannesModel1. The settings for the state modeling unmethylated regions were directly transferred.

The time and memory complexity for analyzing a methylation profile of length T by a MeDIP-HMM of order L during one training step is given by $O(T \cdot N^{L+1})$ with respect to $N = 3$ states. For each of the T measurements all possible N^{L+1} state-transitions must be considered. Thus, the complexity is mainly dominated by the extended state-transition process involving the last L predecessor states to determine the next state (e.g. Figure S3). This time and memory complexity can also be transferred to the mixture models ($L = 0$) and the first-order HMMs ($L = 1$) in Table 1.

To quantify the potential of genomic regions of being methylated, corresponding scores based on state-posterior probabilities were provided by each of the different methods. A score close to one indicates that the corresponding genomic region is a potential candidate of being methylated, while a score close to zero specifies that this region is potentially unmethylated. For methods that do not allow the simultaneous analysis of both replicates of our root methylome data set, average state-posterior probabilities obtained from separate analyses of replicates were considered as scores.

Based on these scores, the identification of candidate regions of methylation was evaluated for the different methods. For each method, we computed the true-positive-rate (TPR) at a fixed false-positive-rate (FPR) of 1%. To enable a more global performance comparison, we also computed the area under the receiver-operating-characteristic curve (AU-ROC) and the area under the precision-recall curve (AU-PRC). Because runtimes of methods can also be an important for the analysis of high-density tiling array data, we additionally measured the time required by each method for performing the analysis of our root methylome data. All results are summarized in Table 1. Corresponding ROC and PRC curves are shown in Figure S4. An additional summary for MeDIP-HMMs of order zero up to four is given in Table S3.

The best performing methods in terms of accuracy of identifying methylated regions are the proposed MeDIP-HMM, CortijoHMM (Cortijo *et al.* (2012)) and tileHMM (Humburg *et al.* (2008)). In comparison to CortijoHMM, which has been specifically developed for the analysis of Arabidopsis MeDIP-chip data, MeDIP-HMM reaches the same performance without being dependent on organism-specific settings. As expected, the second-order MeDIP-HMM required longer for performing the analysis than the first-order CortijoHMM. The first-order MeDIP-HMM reaches nearly identical AU-ROC and AU-PRC values, but could not reach the level of TPR at 1% FPR as CortijoHMM (Table S3). Thus, at the price of a slightly higher runtime, the second-order MeDIP-HMM is able to compensate Arabidopsis-specific ad hoc settings required by CortijoHMM. Compared to tileHMM, MeDIP-HMM reaches a higher TPR at 1% FPR. Additionally, MeDIP-HMM is faster than tileHMM and provides the possibility to differentiate between different methylation levels due to the usage of three states. StandardHMM is also reaching a good performance, but having a smaller TPR at 1% FPR compared to MeDIP-HMM, CortijoHMM and tileHMM. Generally, all these four methods reach comparable global performances as indicated by nearly identical AU-ROC and AU-PRC values.

Comparing all six tested HMM-based methods, TileMAPHMM (Ji and Wong (2005)) reaches the lowest performance potentially because this method does not use a training algorithm, which also leads to the fastest runtime among all tested methods. More generally, MeDIP-HMM, CortijoHMM, tileHMM, StandardHMM and ChIPmixHMM are reaching clearly higher TPRs at 1% FPR and higher AU-PRCs than methods based on mixture models. This is again obtained due to the modeling of dependencies between measurements of directly adjacent chromosomal regions, which cannot be realized by mixture models. Further support to this is given considering all ChIPmix-based methods (Martin-Magniette *et al.* (2008)) for which ChIPmixHMM clearly outperforms multiChIPmix and ChIPmix that are both utilizing a mixture model. The best method based on a mixture model is JohannesModel1 reaching an accuracy comparable to that of TileMAPHMM, but requiring nearly eighty-one times longer for the analysis.

Generally, also these comparisons indicate that the second-order MeDIP-HMM is well-suited for the identification of methylated regions. This model reaches a high accuracy, has a low runtime, does not depend on organism-specific settings and additionally enables to differentiate between different levels of DNA methylation. Additionally, considering more stringent validation data from Zilberman *et al.* (2008) than utilized for this comparison, the second-order MeDIP-HMM is clearly outperforming all other tested methods (Figure S4).

3.3 Functional analysis of DNA methylation states identified by MeDIP-HMM

The Arabidopsis Information Resource (TAIR8) genome annotation (Rhee *et al.* (2003)) enables us to analyze whether predictions by the second-order MeDIP-HMM are targeting specific functional units of the genome and to which extent these findings are in accordance with current knowledge on the Arabidopsis methylome. We initially assigned each genomic region to its most likely underlying state (unmethylated, methylated, or highly methylated) of the MeDIP-HMM using state-posterior decodings. This resulted in $463,877$ (65.2%) unmethylated, $143,017$ (20.1%) methylated, and $104,426$ (14.7%) highly methylated regions. Thus, 65.2% of regions are unmethylated, whereas 34.8% of regions are targeted by methylation, which is in good agreement with previous studies for aerial and root tissues by Zilberman *et al.* (2007, 2008).

We next classified all regions according to the genome annotation and identified under- or over-representations of specific functional categories by sampling the same numbers of unmethylated, methylated, and highly methylated regions randomly from all regions on the tiling array using 500 repeats (Figure 3).

Regions identified as being unmethylated are enriched in genic categories (gene, mRNA, protein, exon, CDS), and 5' and 3' untranslated regions (UTRs), whereas transposons are clearly under-represented (Figure 3a). About 62.4% of all genes and 21.2% of all transposons are identified as being unmethylated over their whole tiled sequences (Figure S5). This is in good agreement with results of previous studies (e.g. Zilberman *et al.* (2007); Cokus *et al.* (2008); Ahmed *et al.* (2011)). Furthermore, using the chromatin classification by Bernatavichute *et al.* (2008), we find that 72.8% of unmethylated regions are located in euchromatin, which is in agreement with these regions being gene-rich and transposon-poor (Table S4).

Regions identified as being methylated include a large fraction of genic regions and a small fraction of 5' and 3' untranslated regions (Figure 3b). About 35.6% of genes are at least partially methylated (Figure S5). This is in good accordance with previous findings of gene body methylation (Zhang *et al.* (2006); Zilberman *et al.* (2007); Cokus *et al.* (2008)). Additionally, regions identified as being methylated are clearly enriched in transposons. This is expected and coincides with the three previous studies, because methylation of transposons is one known mechanism to silence these elements (Law and Jacobsen (2010); Teixeira and Colot (2010)). Overall, the fraction of methylated regions modeled by state $'\mathrm{I}'$ targeting genic regions is greater than that of transposons. Generally, the prevalence of methylated regions in euchromatin is significantly enriched comprising about 53% of methylated regions (Table S4), which is much less stronger than observed

Table 1. Performance comparison of different methods applied to the analysis of the MeDIP-chip root methylome data set. The methods are compared based on their identification of methylated DNA-regions in the root methylome data set. The 'Method' column contains the shortcuts of the different methods. The 'Model' column specifies the basic model of the corresponding method. This is either a Hidden Markov Model (HMM) or a Mixture Model (MixMod). The 'SIM' column specifies if a method considers all replicates of an experiment simultaneously. The 'Training' column specifies the algorithm used for adapting the corresponding method to the data. For an HMM-based method this is either a Bayesian Baum-Welch training (Bayesian BW), Viterbi training (Viterbi) or a standard Baum-Welch training (BW). For methods based on a mixture model specific versions of the Expectation Maximization (EM) algorithm are used. The 'Reference' column provides the link for getting more information about a specific method. The methods are compared based on different criteria considering the true-positive-rate (TPR) reached at a fixed false-positive-rate (FPR) of 1%, the area under the receiver-operating-characteristic curve (AU-ROC), the area under the precision-recall curve (AU-PRC) and the runtime in seconds required for the complete analysis of the data set. The runtime was measured on a standard desktop computer with 2.6 GHz and 4 GB of memory, except for CortijoHMM evaluated on a cluster node with 3 GHz and 8 GB of memory. For MeDIP-HMM only results obtained by the second-order model are shown. Corresponding ROC and PRC curves of all methods are shown in Figure S4. An additional summary for MeDIP-HMMs of order zero up to four is given in Table S3.

Method	Model	SIM	Training	Reference	TPR at 1% FPR	AU-ROC	AU-PRC	Runtime
MeDIP-HMM	HMM	yes	Bayesian BW	this manuscript	0.66	0.98	0.93	637s
CortijoHMM	HMM	no	Constrained BW	Cortijo *et al.* (2012)	0.66	0.98	0.93	402s
tileHMM	HMM	no	Viterbi + BW	Humburg *et al.* (2008)	0.65	0.98	0.93	741s
StandardHMM	HMM	no	Bayesian BW	Seifert *et al.* (2009)	0.63	0.98	0.93	218s
ChIPmixHMM	HMM	no	BW	C. Bérard (pers. comm.)*	0.58	0.97	0.91	914s
TileMAPHMM	HMM	yes	-	Ji and Wong (2005)	0.56	0.96	0.90	61s
JohannesModel1	MixMod	yes	Incremental EM	Johannes *et al.* (2010)	0.55	0.97	0.90	4936s
multiChIPmix	MixMod	yes	EM	C. Bérard (pers. comm.)*	0.52	0.97	0.90	1109s
ChIPmix	MixMod	no	EM	Martin-Magniette *et al.* (2008)	0.51	0.97	0.90	1204s

* Source code available upon request from C. Bérard (caroline.berard@agroparistech.fr)

for unmethylated regions because transposons identified as being methylated are mostly located within heterochromatic regions.

Regions identified as being highly methylated are clearly enriched in transposons, whereas only very small fractions of genic regions, 5' and 3' untranslated regions are highly methylated (Figure 3c). About 27.2% of all transposons and only about 2% of all genes are highly methylated over their whole tiled sequences (Figure S5). This coincides with previous findings (Zilberman *et al.* (2008); Ahmed *et al.* (2011)). Overall, we find that highly methylated regions modeled by state $'M'$ are preferentially located in heterochromatin. We identified a significantly enriched proportion of 86.3% heterochromatic regions compared to only 13.7% euchromatic regions (Table S4).

We further analyzed identified DNA methylation states of transposon superfamilies utilizing the extended TAIR8 transposon annotation of Ahmed *et al.* (2011). We observe clear tendencies that especially unmethylated or methylated transposons of the superfamilies LTR/Gypsy, LTR/Copia, LINE/L1 and DNA/En-Spm are much smaller in size than highly methylated members of the same superfamily (Figure S6). Shorter sizes and less strong methylation are known to be indicative of non-functional transposon relics (Ahmed *et al.* (2011)). These dependencies can only be observed so explicitly using the three-state decoding of MeDIP-HMM.

Summing up, predictions by MeDIP-HMM are in good accordance with current knowledge on the Arabidopsis methylome assembled in previous studies using tiling array and bisulfite sequencing experiments. Moreover, differentiations between unmethylated, methylated and highly methylated regions are only possible with MeDIP-HMM and cannot be achieved using standard methods for ChIP-chip analysis designed to discriminate between two populations of measurements. Thus, MeDIP-HMM allows a more detailed analysis and potentially a more advanced interpretation of results.

3.4 Application of MeDIP-HMM to non-Arabidopsis data

To illustrate the general applicability of MeDIP-HMM, we analyzed publicly available promoter DNA methylation data of brain tissue of domesticated White Leghorn (WL) and wild type Red Junglefowl (RJF) chickens (Nätt *et al.* (2012)). We downloaded processed data for parents and offspring chickens from ArrayExpress (E-MTAB-648, E-MTAB-649) and created a small data set containing measurements of tiles in corresponding promoter regions for two selected genes ABHD7 and PCDHAC1 shown to have a stable transgenerational methylation profile. This data set includes four replicates (females and males each with high and low fear behavior) for WL and RJF. We trained a second-order MeDIP-HMM with four-variate Gaussian emissions on this data set using initial proportions of unmethylated and highly methylated tiles like applied for the Arabidopsis root methylome data. Finally, we performed a state-posterior decoding into the most likely underlying methylation states. The results are shown in Figure S7. In accordance with Nätt *et al.* (2012), we also find that the methylation patterns of ABHD7 and PCDHAC1 are very stable between parents and offspring. Additionally, the decoding into three states by MeDIP-HMM enables a more detailed view on promoter DNA methylation potentially highlighting domains with high and low levels of DNA methylation. Thus, this additional study indicates that MeDIP-HMM can also be applied to non-plant data.

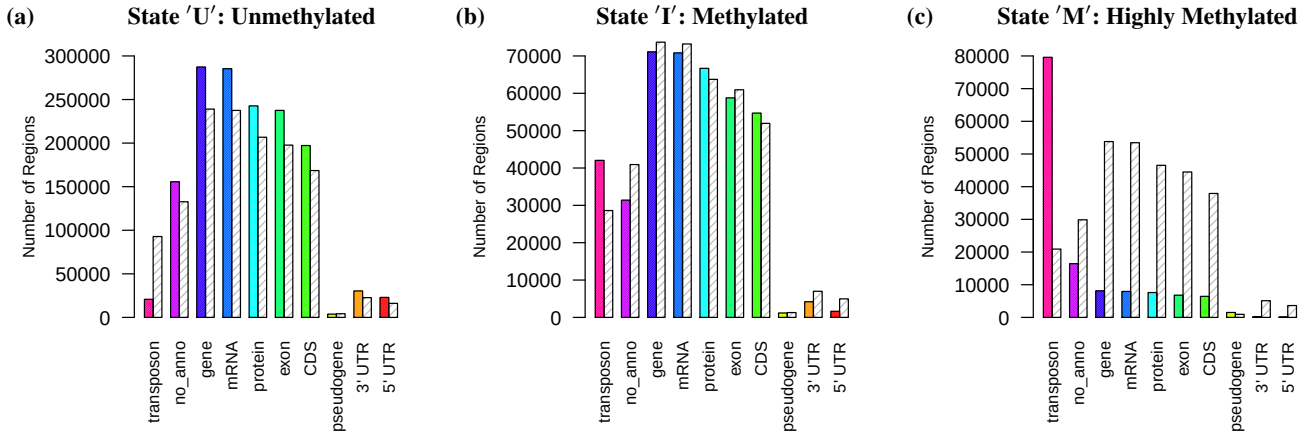

Fig. 3. Functional analysis of genomic regions in the Arabidopsis root methylome data set according to underlying DNA methylation states identified by MeDIP-HMM. Genomic regions were annotated using the functional categories defined by the TAIR8 genome annotation. Colored bars represent the number of regions in each category identified by MeDIP-HMM. Grey shaded bars represent the average number of regions obtained for each category by randomly sampling the same number of regions from all genomic regions in the data set using 500 repeats. Genomic regions identified as being unmethylated are shown in **(a)**, regions identified as being methylated are shown in **(b)**, and regions identified as being highly methylated are shown in **(c)**. All identified genomic regions classified into the different categories are significantly different from random sampling with p-values less than 0.01. The states 'I' and 'M' modeling different degrees of DNA methylation are having obviously distinct functional interpretations. Less strongly methylated transposons and genes with gene body methylation are covered by state 'I', whereas strongly methylated transposons are mainly characterizing regions assigned to state 'M'.

4 CONCLUSIONS

We developed a three-state MeDIP-HMM for the analysis of DNA methylation data from high-density tiling arrays enabling the identification of distinct methylation levels. MeDIP-HMM enables a simultaneous analysis of replicates and improves modeling of spatial dependencies between chromosomal regions using a higher-order state-transition process.

We carefully evaluated our model and existing methods for ChIP-chip analyses on DNA methylation data of Arabidopsis root tissue. Compared to these methods, MeDIP-HMM reached an overall good performance in combination with a fast runtime. This also revealed that HMMs modeling spatial dependencies between chromosomal regions are generally better suited for the analysis of MeDIP-chip data than mixture models ignoring these dependencies. Besides this, moderately higher-order MeDIP-HMMs were identified to be more precise than the first-order MeDIP-HMM and other existing first-order HMMs. This is in good accordance with a previous study in comparative genomics (Seifert *et al.* (2012)). Generally, we identified that a second-order MeDIP-HMM is working best on the root methylome data reaching a performance comparable to a three-state first-order HMM specifically developed for the analysis Arabidopsis MeDIP-chip data in a companion study by Cortijo *et al.* (2012). The companion approach utilizes the observation that introns in Arabidopsis are typically found to be unmethylated for estimating the emission parameters of the unmethylated state. This cannot be directly transferred to other organisms like human, where introns are usually found to be methylated (Lister *et al.* (2009)). Our results indicate that the second-order MeDIP-HMM is able to compensate Arabidopsis-specific ad hoc settings required by the companion approach. Moreover, we also showed that our MeDIP-HMM can be applied to the analysis of non-plant data as well. Overall, MeDIP-HMM is more versatile by being independent of organism-specific settings, utilizing a higher-order state-transition process, and enabling simultaneous analyses of replicates.

Generally, the differentiation between unmethylated, methylated, and highly methylated regions by MeDIP-HMM enables an improved interpretation of predictions compared to existing ChIP-chip analyses methods only enabling a binary classification into unmethylated and methylated regions. Moreover, the predictions of MeDIP-HMM were in good accordance with current knowledge on the Arabidopsis methylome. All these findings clearly indicate that MeDIP-HMM is a useful method for the analysis of DNA methylation data. Given that MeDIP-HMM is independent of organism-specific settings, it should be applicable to the analysis of the methylomes of plant and non-plant species equally.

ACKNOWLEDGEMENTS

We thank Caroline Bérard (AgroParisTech) for providing source codes of multiChIPmix and ChIPmixHMM. We thank Jens Keilwagen (IPK) and Jan Grau (University of Halle) for Jstacs support. We thank Marc Strickert (University of Marburg) for critical reading of the manuscript. We thank the anonymous reviewers for their valuable comments. This work was supported by the Ministry of Culture Saxony-Anhalt (grant XP3624HP/0606T), DAAD PROCOPE (grant 50748812), EU FP7 Network of Excellence 'EPIGENESYS' and ANR-GPLA 'REGENEOME'.

REFERENCES

Ahmed, I., Sarazin, A., *et al.* (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Res*, **39**, 6919–6931.

Banaei, A. M., Roudier, F., *et al.* (2011). Additive inheritance of histone modifications in Arabidopsis thaliana intraspecific hybrids. *Plant J*, **67**, 691 – 700.

Barlow, D. P. (2011). Genomic Imprinting: A Mammalian Epigenetic Discovery Model. *Annu Rev Genet*, **45**, 379–403.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Beck, S. and Rakyan, V. K. (2008). The methylome: approaches for global DNA methylation profiling. *Trends Genet*, **24**, 231–236.

Berchtold, A. and Raftery, A. E. (2002). The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Stat. Sci.*, **17**, 328–356.

Bernatavichute, Y. V., Zhang, X., *et al.* (2008). Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE*, **3**(9), e3156.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its applications to parameter estimation for Gaussian mixture and Hidden Markov Models. *Technical Report ICSI-TR 97-021*.

Bolstad, B. M., Irizarry, R. A., *et al.* (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Borgel, J., Guibert, S., *et al.* (2010). Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet*, **42**, 1093–1100.

Cokus, S. J., Feng, S., *et al.* (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**(6), 215–219.

Cortijo, S., Wardenaar, R., *et al.* (2012). *Genome-wide analysis of DNA methylation in Arabidopsis using MeDIP-chip*. Plant Epigenome: Understanding and Analysis. Humana Press/Springer: New York, in press.

Dempster, A. P., Laird, N. M., *et al.* (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

du Preez, J. A. (1998). Efficient training of high-order hidden Markov models using first-order representations. *Comput Speech Lang*, **12**, 23–39.

Durbin, R., Eddy, S., *et al.* (1998). *Biological sequence analysis - Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eng, C., Asthana, C., *et al.* (2009). A New Data Mining Approach for the Detection of Bacterial Promoters Combining Stochastic and Combinatorial Methods. *J. Comp. Biol.*, **16**, 1211–1225.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modifications. *Nat Rev Genet*, **8**, 286–298.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum A Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. Speech Audio Process.*, **2**, 291–298.

Harrison, A. and Parle-McDermott, A. (2011). DNA methylation: a timeline of methods and applications. *Front Genet*, **2**(74), 1–13.

Humburg, P., Bulger, D., *et al.* (2008). Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics*, **9**(2).

Ji, H. and Wong, W. H. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.

Johannes, F., Wardenaar, R., *et al.* (2010). Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics*, **26**, 1000–1006.

Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*, **11**, 204–220.

Li, W., Meyer, C. A., *et al.* (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**, i274–i282.

Lister, R., O'Malley, R. C., *et al.* (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Lister, R., M., P., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **426**, 315–322.

Mari, J.-F., Halton, J.-P., *et al.* (1997). Automatic word recognition based on second-order hidden Markov models. *IEEE T. Speech and Audio P.*, **5**, 22–25.

Martin-Magniette, M.-L., Mary-Huard, T., *et al.* (2008). ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, **24**, i181–i186.

Nätt, D., Rubin, C.-J., *et al.* (2012). Heritable genome-wide variation of gene expression and promoter methylation between wild and domesticated chickens. *BMC Genomics*, **13**(59).

Rhee, S. Y., Beavis, W., *et al.* (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, **31**, 224–228.

Seifert, M. (2010). *Extensions of Hidden Markov Models for the analysis of DNA microarray data*. Ph.D. thesis, Martin Luther University Halle.

Seifert, M., Keilwagen, J., *et al.* (2009). Utilizing gene pair orientations for HMM-based analysis of ChIP-chip data. *Bioinformatics*, **25**, 2118–2125.

Seifert, M., Strickert, M., *et al.* (2011). Exploiting prior knowledge and gene distances in the analysis of tumor expression profiles with extended Hidden Markov Models. *Bioinformatics*, **27**, 1645–1652.

Seifert, M., Gohr, A., *et al.* (2012). Parsimonious Higher-Order Hidden Markov Models for Improved Array-CGH Analysis with Applications to Arabidopsis thaliana. *PLoS Comp Biol*, **8**(1), e1002286.

Teixeira, F. K. and Colot, V. (2010). Repeat elements and the Arabidopsis DNA methylation landscape. *Heredity*, **105**, 14–23.

Wang, Y. (2006). The Variable-length Hidden Markov Model and Its Applications on Sequential Data Mining. *Technical Report, Department of Computer Science, Tsinghua University, Beijing, China.*

Wutz, A. (2011). Gene silencing X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat Rev Genet*, **12**, 542–553.

Zhang, X., Yazaki, J., *et al.* (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**, 1189–1201.

Zilberman, D., Gehring, M., *et al.* (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, **39**, 61–69.

Zilberman, D., Coleman-Derr, D., *et al.* (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, **456**(6), 125–130.