Nucleic Acids Research, 2016 **1–9**
doi: 10.1093/nar/gkw1319

HMCan-diff: a method to detect changes in histone modifications in cells with different genetic characteristics

Haitham Ashoor[1], Caroline Louis-Brennetot[2], Isabelle Janoueix-Lerosey[2], Vladimir B. Bajic[1,*] and Valentina Boeva[3,4,*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal 23955-6900, Saudi Arabia, [2]Institut Curie, Inserm U830, PSL Research University, F-75005, Paris, France, [3]Institut Curie, Inserm U900, Mines ParisTech, PSL Research University, F-75005 Paris, France and [4]Institut Cochin, Inserm U1016, CNRS UMR 8104, Université Paris Descartes UMR-S1016, F-75014 Paris, France

Received April 01, 2016; Revised December 08, 2016; Editorial Decision December 15, 2016; Accepted December 19, 2016

ABSTRACT

Comparing histone modification profiles between cancer and normal states, or across different tumor samples, can provide insights into understanding cancer initiation, progression and response to therapy. ChIP-seq histone modification data of cancer samples are distorted by copy number variation innate to any cancer cell. We present HMCan-diff, the first method designed to analyze ChIP-seq data to detect changes in histone modifications between two cancer samples of different genetic backgrounds, or between a cancer sample and a normal control. HMCan-diff explicitly corrects for copy number bias, and for other biases in the ChIP-seq data, which significantly improves prediction accuracy compared to methods that do not consider such corrections. On *in silico* simulated ChIP-seq data generated using genomes with differences in copy number profiles, HMCan-diff shows a much better performance compared to other methods that have no correction for copy number bias. Additionally, we benchmarked HMCan-diff on four experimental datasets, characterizing two histone marks in two different scenarios. We correlated changes in histone modifications between a cancer and a normal control sample with changes in gene expression. On all experimental datasets, HMCan-diff demonstrated better performance compared to the other methods.

INTRODUCTION

The development of ChIP-seq technology (1) has enabled the construction of genome-wide maps of protein–DNA interactions. Such maps provide information about transcriptional regulation at the epigenetic level (histone modifications and histone variants) and at the level of transcription factor activity. Recently, thousands of ChIP-seq datasets have been produced by different consortia including ENCODE (2) and the NIH Roadmap Epigenomics Mapping Consortium (3). The data produced contain histone modification libraries for both normal and cancer cell karyotypes.

In cancer, genetic and epigenetic abnormalities cooperate in the process of regulating activities of oncogenes and onco-suppressors (4). For example, lower levels of trimethylation of lysine 36 of histone H3 (H3K36me3) and trimethylation of lysine 20 of histone H4 (H4K20me3) in proximity of the gene *NSD1,* contribute to the development of nervous system tumors (5). Also, higher levels of trimethylation of lysine 27 of histone H3 (H3K27me3), in proximity to the *HOX* cluster of genes, plays a role in prostate cancer (6). Given the role of histone modifications and other epigenetic modifications in cancer, several epigenetic therapy methods have been proposed (7,8).

To better characterize changes in histone modifications and understand epigenetic mechanisms driving cancer initiation, progression and response to therapy, methods to detect changes in histone modifications between pairs of conditions are needed. The demand to design methods to handle ChIP-seq data from cancer samples has been highlighted in several studies (9–12). This demand rises from the fact that cancer genomes are characterized by copy number aberrations. These can introduce statistical biases in downstream analyses that affect results by introducing false positive and false negative predictions.

*To whom correspondence should be addressed. Valentina Boeva. Tel: +33-1-56-24-69-88; Fax: +33-1-56-24-69-11; Email: valentina.boeva@inserm.fr
Correspondence may also be addressed to Vladimir Bajic. Tel: +966-5447-00088; Fax: +966 12 808-2386; Email: vladimir.bajic@kaust.edu.sa

Many methods have been developed to detect regions that exhibit changes in a ChIP-seq signal between two conditions (differential peaks). Some of these methods have been specifically designed to predict differential peaks from narrow marks, such as DiffBind (13), ChIPComp (14) and DBChIP (15), while other methods, such as ChIPDiff (16), ChIPnorm (17) and RSEG (18), have been designed to detect differential peaks from broad marks. Moreover, some methods for differential peak calling require providing sets of peaks in order to identify differential regions. Examples of these methods include MAnorm (19), DiffBind (13) and DBChIP (15). Other methods, such as ODIN (20), MEDIPS (21) and PePr (22), do not require peak regions as an input and are expected to perform equally well for narrow and broad histone marks. Moreover, some methods can account for experiments with either biological or technical replicates (PePr (22), DiffBind (13) and csaw (23)), while other methods cannot (ODIN (20), ChIPDiff (16) and MACS2).

In this study, we introduce HMCan-diff, a method for identifying changes in histone modifications from ChIP-seq cancer data. Our method corrects for copy number aberrations, GC-content bias, sequencing depth, mappability, and noise level, thus accounting for different technical artifacts of ChIP-seq data, and utilizes information from replicates to reduce technical variation effects.

We compared HMCan-diff with several recent and most commonly-used methods, namely ChIPDiff (16), MAnorm (19), MEDIPS (21), ODIN (20), MACS2 (https://github.com/taoliu/MACS/tree/master/MACS2), DiffBind (13), RSEG (18) and csaw (23). We conducted experiments on both simulated and experimental data. On simulated data containing copy number bias, HMCan-diff showed significant performance improvement compared to other tools. HMCan also showed comparable performance on simulated data without copy number bias. On experimental data, HMCan-diff predicted differential histone modification regions that correlate better with changes in gene expression compared to the predictions obtained by other methods, suggesting it has higher accuracy.

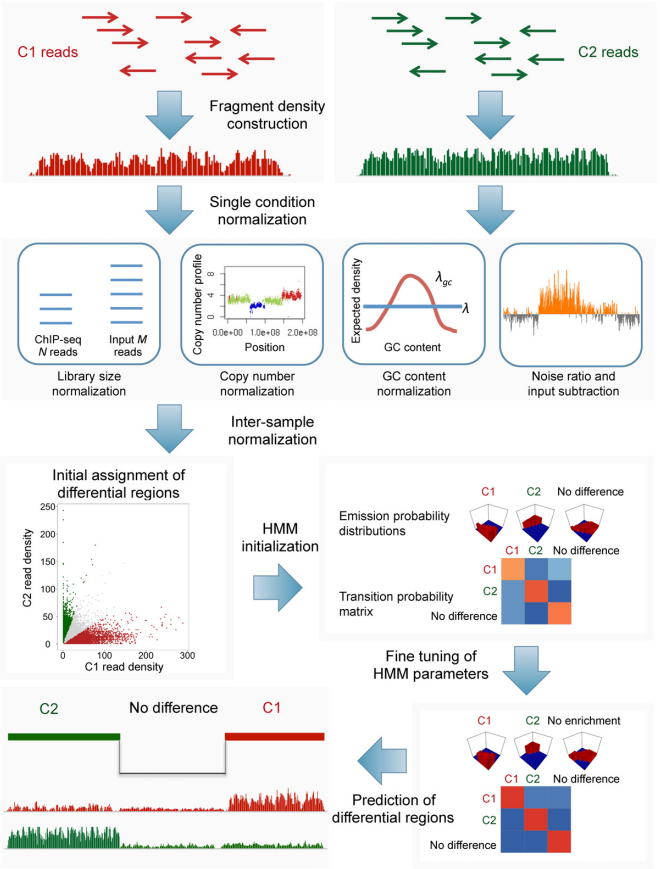

Figure 1. A workflow illustrating HMCan-diff steps. Initially, HMCan-diff constructs a fragment density profile for each provided ChIP-seq or input dataset. Then, it normalizes density profiles of each replicate in each condition for several types of bias, specifically for copy number variation, library size, GC-content bias and noise level. After that, HMCan-diff conducts additional normalization to eliminate further technical variation between conditions. It initializes HMM parameters based on the data. In particular, HMCan-diff defines the HMM emission probability distribution as the joint empirical distribution of normalized density values. Then, HMCan-diff improves these parameters using the Baum-Welch algorithm, and finishes by dividing genomic regions into three states: *C1* (enriched in condition 1), *C2* (enriched in condition 2), and the 'no difference' state.

MATERIALS AND METHODS

Description of HMCan-diff

The HMCan-diff workflow consists of several steps (Figure 1): (i) construction of normalized ChIP-seq density, (ii) inter-conditional normalization, (iii) initialization of the hidden Markov model (HMM) and (iv) learning of HMM parameters and identification of differential peaks. HMCan-diff implements a 3-state multivariate HMM to identify changes in histone modifications; the states are: 'enriched in condition 1′ (C1), 'enriched in condition 2′ (C2), and a 'no difference' state. HMCan-diff is implemented in C++ and is available at http://www.cbrc.kaust.edu.sa/hmcan/.

Construction of normalized density profiles

HMCan-diff uses density construction and normalization methods implemented in the HMCan algorithm (10). The normalization steps include normalization for copy number variation, library size, mappability, GC-content bias, and noise level.

Density profile construction. HMCan-diff transforms reads of ChIP and control samples into fragment density profiles. To do so, HMCan-diff extends reads to the length of initial DNA fragments using the triangular distribution first proposed in the FindPeaks peak calling method (24). In this way, HMCan-diff accounts for variable fragment lengths present in a ChIP-seq experiment. Minimum, median and maximum fragment lengths used to define the triangular distribution can be obtained from the sequencing platform. After raw density construction, HMCan-diff reports density values every N bp, where N is a step fixed by the user. If they like, users can choose the step size as $minL/2$, where $minL$ is the minimum fragment length. Smaller step sizes will not increase the performance but will

provide a better profile resolution for further downstream analysis. In this study we set N to be 50 bp.

HMCan-diff also proposes to mask densities from genomic regions known to contain false signals. HMCan-diff provides blacklisted regions recommended by the ENCODE consortium (2) as its default set of regions.

Correction for copy number alterations. HMCan-diff utilizes the Control-FREEC algorithm (25,26) to learn the copy number profile from input DNA data (control sample). Control-FREEC partitions chromosomes into large genomic windows (in this study we used 100 kb), then fits a polynomial function to model the relationship between the read count per window and GC-content values. Control-FREEC uses this function to normalize the read count for GC-content bias and produce a smooth profile of corrected read counts. Then, it uses a LASSO-based approach to segment the normalized copy number profile. For each segment, Control-FREEC provides the median value of the normalized read count (*medNRC*). Copy neutral regions will correspond to *medNRC* values close to 1; a region deleted in one copy out of 2 will have a *medNRC* value of around 0.5; a region duplicated in a diploid genome will have a *medNRC* value of ∼1.5, and so on. After the ChIP-seq density profile is constructed, HMCan-diff divides density values by *medNRCs* of corresponding segments. In this way, HMCan-diff normalizes both ChIP and input densities for copy number alterations.

Library size normalization. HMCan adjusts the ChIP density values for the library size. Given that the ChIP sample has N_1 reads and the control sample N_2, HMCan-diff scales ChIP density by a factor of (N_2/N_1).

Identifying enriched density regions. Identifying enriched density regions is important for GC-content bias correction and HMM initialization steps. HMCan-diff provides two options to decide whether a given density value corresponds to an enriched region or background region. The first option is to conduct a one-sided Poisson exact test using a local Poisson distribution, where the mean parameter for the Poisson distribution corresponds to the control density value of that position. The second option is to perform the test using the negative binomial (NB) distribution to identify enriched regions. HMCan-diff learns parameters of the NB distribution in a similar way to the method proposed in (27). HMCan-diff sets the first option as its default for identifying enriched density regions.

GC-content normalization. It has been reported that GC-content affects the sequencing depth of Illumina reads (28). Therefore, HMCan-diff implements a correction of ChIP-seq density data for GC-content bias.

First, HMCan-diff associates each density value with a GC-content value. For each density value, HMCan-diff calculates the GC-content for a window of length equal to twice the median fragment length centered on that value, and assigns it to this density value. Then, HMCan-diff groups GC-values into different strata (groups). For example, the first group will have GC content values from the interval [0, 0.20), the second will have values from the interval [0.20, 0.22), the third from the interval [0.22, 0.23), etc. For each value gc in the strata, D_{gc} defines the total sum of density values associated with a given GC-content gc, and N_{gc} the total number of windows that have GC-content gc. The expected density for a specific gc value (λ_{gc}) is calculated as:

$$\lambda_{gc} = \frac{D_{gc}}{N_{gc}} \tag{1}$$

The average expected density over the genome, λ, is calculated as:

$$\lambda = \frac{\sum_{gc} D_{gc}}{\sum_{gc} N_{gc}} \tag{2}$$

HMCan-diff corrects a density value d coming from a region with GC-content gc as:

$$d_{corrected} = \frac{d \cdot \lambda}{\lambda_{gc}} \tag{3}$$

HMCan-diff corrects ChIP-seq data and input control samples independently. Correcting both samples independently is more accurate than using information from the input control sample only.

For the calculation of λ_{gc} and λ for the input data, we use all density values. For ChIP-seq data, we need to limit the calculation of λ_{gc} and λ to background regions only, since a high ChIP-seq signal may correspond to GC-rich regions and thus may be considered a GC-bias. Therefore, for the ChIP-seq signal, we first identify enriched signal regions in order to exclude enriched density loci from the calculations. Then, HMCan-diff calculates λ_{gc} and λ from bins corresponding to background regions.

Background subtraction. Since ChIP-seq density values are a mixture of real signal and noise, while input control density values correspond only to noise, we need to rescale the control density to match noise levels in both samples. To do so, HMCan-diff calculates the noise level (λ_{noise}) as:

$$\lambda_{noise} = \lambda_{ChIP} / \lambda_{control} \tag{4}$$

Lastly, HMCan-diff calculates the final corrected density profile as:

$$d_{final} = d_{ChIP\ Corrected} - d_{Control\ corrected} \cdot \lambda_{noise} \tag{5}$$

HMCan-diff applies the above-described normalization steps to each ChIP-seq sample for each condition, independently.

These steps result in the creation of a normalized density profile d_{mij} for each position m (every N bp) of sample S_{ij}, where i is the condition and j the replicate.

Inter-sample normalization

The basic idea behind inter-sample normalization in HMCan-diff is to adjust the total normalized density in all samples to similar levels. In order to achieve this, HMCan-diff calculates the total genome-wide density in each replicate per condition, X_{ij}:

$$X_{ij} = \sum_m d_{mij} \tag{6}$$

We define the reference sample, *ref*, as the sample with the minimum noise level among all samples to be analyzed. Then, we define normalizing coefficients β_{ij} such that:

$$\boldsymbol{\beta_{ij}} = \boldsymbol{X_{ref}}/\boldsymbol{X_{ij}} \quad (7)$$

Last, we derive the normalized density *d'* as:

$$\boldsymbol{d'_{mij}} = \boldsymbol{\beta_{ij}} \cdot \boldsymbol{d_{mij}} \quad (8)$$

Inter-sample normalization turns out to be very important as it adjusts for different antibody efficiency between ChIP-seq samples. Moreover, inter-sample normalization does not take into account copy number status, as our method has already corrected for that in the previous step.

HMM definition and initialization

A multivariate HMM H is defined by a set $H = \{S, O, T, E\}$, where:

- S is the set of possible states, $S = \{C1, C2, \text{'}no\ difference\text{'}\}$
- is the set of observations; we define O in HMCan-diff as the vector of normalized densities d'_{mij} for each genomic position, from all replicates and conditions
- T is the set of transition probabilities
- E is the set of emission probabilities derived from the empirical joint distribution of the data.

Initially, to simplify our model and avoid bias towards the '*no difference*' state, we discard all non-enriched regions from subsequent calculations (see 'Identifying enriched regions' section). We discard loci that the test indicates as non-enriched in all replicates, from both conditions. For the remaining loci, we use the fold change to initialize HMM emission and transition probabilities, where for each position m, we define the fold change fc_m as:

$$\boldsymbol{fc_m} = \textbf{median}\left\{\frac{\boldsymbol{d'_{m1i+1}}}{\boldsymbol{d'_{m2j+1}}}; \boldsymbol{i} \in \overline{1:k}; \boldsymbol{j} \in \overline{1:l}\right\}, \quad (9)$$

where k and l are the number of replicates in conditions 1 and 2. Next, we initialize the HMM density value state as follows:

- If $(fc_m > T) \rightarrow C1$ state;
- If $(fc_m < 1/T) \rightarrow C2$ state;
- Otherwise $\rightarrow$ '*no difference*' state.

Here, T is a threshold for median fold change, set by the user (default value is 2).

Learning HMM parameters and identifying differential regions

We use the Baum-Welch algorithm (29) to learn emission and transition probabilities. Thereafter, we use threshold-based posterior decoding to decode the final sequence of states. After mapping each density value into its corresponding state, neighboring bins possessing the same differential state (either *C1* or *C2*) are merged to compose differential peaks.

For each differential peak we calculate peak score (PS) using the log likelihood ratio:

$$PS = log\frac{P(differential\ state|region)}{P(\neg differential\ state|region)}, \quad (10)$$

where a differential state could be a *C1* or *C2* one. Higher values of peak scores correspond to higher confidence regions in HMCan-diff predictions.

Two post-processing steps follow calling the differential peaks: differential peaks with a length of less than the median fragment length are ignored, and differential peaks showing the same state (*C1* or *C2*) within a distance below a user-defined value are merged into a single region (default distance 1 kb).

HMCan-diff provides outputs for differential peaks in the standard BED format. It creates two files: one for narrow peaks and another for broad peaks (regions). It also provides normalized density WIG files, which can be helpful for the visual inspection of data as well as for downstream analyses.

Simulated data

We conducted two simulation experiments. In both, we simulated a hypothetical histone mark across conditions 1 and 2, and along human chromosome 1 only (from the hg19 assembly). We set the fragment length to 150 bp and read length to 76 bp. We simulated the ChIP-seq histone modification region's length randomly from 1 to 20 kb. The simulated signal covered 10% of chromosome 1 in each condition, and differential regions represented 25% of the simulated signal. For each condition, we simulated three replicates. We simulated the technical variation between the replicates by controlling the noise level and GC-content bias in each replicate. For the noise level, we simulated its values by picking a random value from a normal distribution $N \sim (\mu, \sigma)$, (μ denoting the mean, σ the standard deviation), where we set $\mu_1 = 0.75$ for condition 1 and $\mu_2 = 0.5$ for condition 2, with $\sigma = 0.1$ for both. We use a 'noise level' parameter to control fold change between signal and background; a higher noise level corresponds to lower antibody efficiency. In all our simulations, we set minimal fold change between signal and background equal to 2. To simulate GC-content bias, we used six different GC-content bias profiles from different ENCODE datasets (see Supplementary Tables S1 and S2 for more details on the parameters for simulated data).

In the first simulation (simulation 1), all regions had a normal copy number, while in the second (simulation 2) we simulated differences in copy number between conditions. Reads were simulated using a ChIP-seq data simulation tool that accompanied the HMCan method (10). To simulate differential regions in simulation 1, we controlled the ratio of ChIP-seq read counts across regions to reflect the state of the simulated regions (*C1*, *C2* or '*no difference*'); we considered a region to be in a differential state when its read ratio was higher than 2. For simulation 2, first we divided chr1 into 12 segments of equal length. We assigned each segment a different copy number status for each condition (Supplementary Table S3). In the case of signal loci, we assigned a different allele count for signal regions in each segment.

We defined differential regions using the proportion of the signal present in DNA alleles to the actual copy number at that region (i.e., total number of alleles). More precisely, we denoted the number of alleles where the signal is present by A, and the copy number of the region by CN, and defined differential regions as those where $|A_1/CN_1 - A_2/CN_2| >= t$, where we set $t = 0.5$. We used Bowtie (30) with the default parameters to align simulated reads to the reference genomic sequence. The simulated data can be downloaded from http://www.cbrc.kaust.edu.sa/hmcan/Download.php.

Experimental data

We used H3K27me3 and H3K27ac ChIP-seq data, and RNA-seq data, to evaluate the performance of HMCan-diff in two experiments: (i) lung adenocarcinoma (A549 cell line) compared to normal lung tissue; (ii) human breast adenocarcinoma (MCF7 cell line) compared to primary human mammary epithelial cells (HMEC cell line). The A549 data were generated using the Diagenode polyclonal antibody specific to H3K27me3 (C15410069) and the Abcam antibody specific to H3K27ac (ab4729). Data is deposited in GEO with accession number GSE75903. Chromatin preparation and ChIP were performed with the Ideal ChIP-seq kit for histones according to the supplier's protocol (Diagenode). Data for the lung tissue were obtained from the epigenomic roadmap database, and data for the MCF7 and HMEC cell lines were obtained from ENCODE.

RESULTS

Evaluation on simulated data

First, we compared HMCan-diff with other relevant tools: ChIPDiff (16), DiffBind (13), MACS2, MAnorm (19), MEDIPS (21), ODIN (20) and RSEG (18), on simulated data. ChIPDiff uses an HMM with a beta-binomial fixed emission distribution, and learns transition probabilities using the Baum-Welch algorithm (29). DiffBind initially preprocesses ChIP-seq reads that fall into peak regions by subtracting the input read count; it then uses the edgeR package (31) to detect differential regions. MACS2 first calls peaks and constructs read pileup files for both conditions, then determines differential regions by assessing the fold change between conditions. MAnorm constructs an MA-plot from reads falling in the common peak regions between the two conditions, then uses this plot to normalize data between conditions; after normalization it identifies differential peaks. MEDIPS uses statistical methods developed in the edgeR package (31) to identify differential regions from various assays including ChIP-seq. MEDIPS also provides a threshold-based procedure to account for copy number variation in the analyzed data. ODIN uses HMMs to identify differential peaks; it models the emission distribution using a mixture of Poisson distributions. RSEG also uses HMMs; it models emission probabilities as the difference of two independent variables following the NB distribution.

We evaluated the accuracy of each method by constructing precision–recall (PR) curves (32) based on the predictions of each tool. To construct PR curves, we sorted predictions of each algorithm based on their scores or p-values, then for each method we applied different thresholds to get

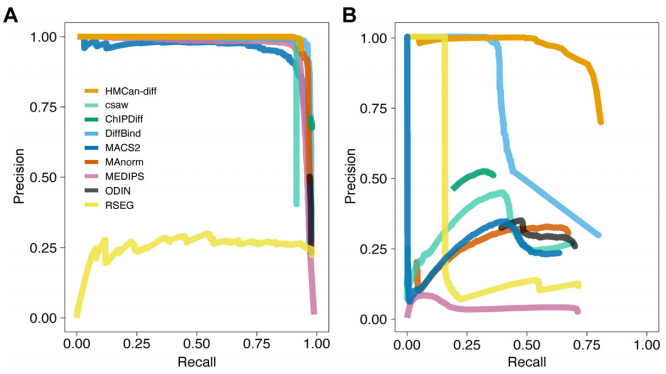

Figure 2. Precision-recall curves for HMCan-diff and other methods on simulated data. (**A**) Precision-recall curves on data simulated without copy number bias: HMCan-diff is slightly better than the majority of tools. (**B**) Precision-recall curves on the simulated data with copy number bias: HMCan-diff shows significantly better performance than the other methods.

different predictions sets. For each threshold value, we considered predictions with score above the threshold as differential. A bin in a region predicted to be differential was considered a true positive (TP) if it overlapped with a simulated differential region; it was considered a false positive (FP) if it overlapped with a non-differential region. Bins from differential regions that were not predicted as such were considered false negatives (FN). Then,

$$Recall = TP/(TP + FN), \quad (11)$$

$$Precision = TP/(TP + FP), \quad (12)$$

where recall quantifies the sensitivity of the method, and precision quantifies the specificity of positive predictions. We reported recall values in regions corresponding to precision values 0.9 and 0.95) and reported precision and recall values at the best cutoff value (Table 1). We defined the best cutoff as that of corresponding to the closest point to the ideal predictor (recall = 1, precision = 1) (33).

We investigated different parameter sets for methods in both simulations (Supplementary Figure S1 and Supplementary Figure S2). Then, we selected parameters that yielded the best performance (Supplementary Table S5). When we compared calculated PR-curves based on data simulated without copy number bias (Figure 2A), we found comparable accuracy of most methods, with a slightly lower value for HMCan-diff compared to DiffBind (F-measure HMCan = 0.95 and DiffBind = 0.96). The poor performance of RSEG in simulation 1 was likely due to not accounting for different antibody efficiencies simulated in our *in silico* experiment.

When we compared PR-curves calculated on a dataset where copy number bias was present (Figure 2B, simulation 2), we noticed a significantly lower performance of tools that did not directly account for copy number bias, unlike HMCan-diff, which maintained good prediction accuracy (Table 1). Additionally, we evaluated the performance of HMCan-diff and other methods on regions having different amplitudes of changes in copy number status between the conditions (Supplementary Figure S3). To do so, we divided regions from simulation 2 into two categories: (i) regions

Table 1. Recall and precision for all methods tested. HMCan-diff shows marginally better performance in simulation 1 (no copy number changes between the two conditions) and significantly better performance in simulation 2 (includes copy number changes between the conditions)

Method	Simulation 1				Simulation 2			
	Precision = 0.95	Precision = 0.90	Best cutoff		Precision = 0.95	Precision = 0.90	Best cutoff	
	Recall	Recall	Recall	Precision	Recall	Recall	Recall	Precision
HMCan-diff	0.94	0.94	0.93	**0.98**	**0.65**	**0.75**	**0.784**	0.875
ChIPDiff	No predictions	No predictions	**0.97**	0.72	No predictions	No predictions	0.36	0.507
csaw	0.92	0.92	0.91	0.41	No predictions	No predictions	0.64	0.27
DiffBind	**0.96**	**0.96**	0.96	0.97	0.38	0.39	0.381	0.96
MACS2	0.82	0.91	0.91	0.9	0.001	0.0001	0.635	0.231
MAnorm	0.93	0.94	0.93	0.95	No predictions	No predictions	0.662	0.317
MEDIPS	0.91	0.93	0.92	0.95	No predictions	No predictions	0.702	0.034
ODIN	No predictions	No predictions	**0.97**	0.5	No predictions	No predictions	0.684	0.26
RSEG	No predictions	No predictions	0.93	0.26	0.16	0.16	0.15	**0.99**

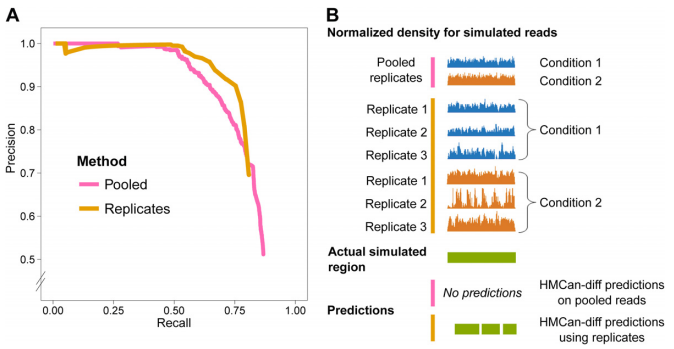

Figure 3. Effects of the use of replicates on HMCan-diff predictions. (**A**) When using replicate information, HMCan-diff produces better predictions than when using pooled data. (**B**) Genome browser view showing that combining data from different replicates may lead to losing the correct differential signal, while the use of information from replicates improves HMCan-diff prediction accuracy.

with a low copy number discrepancy, where the absolute difference in copy number between the two conditions is less than two; and (ii) regions with a high copy number discrepancy, where the copy number difference is greater than or equal to two. In both scenarios, HMCan-diff performed significantly better than other tools. The performance gain in the case of high copy number discrepancy for HMCan-diff was higher than in regions of low copy number discrepancy.

We investigated the effect of applying different combinations of normalizations implemented by HMCan-diff (Supplementary Figure S4). We showed that there is a significant drop in HMCan-diff prediction accuracy when both GC-content and copy number bias normalizations are skipped. When both normalization steps were applied, the recall value of HMCan-diff at a precision of 0.95 was 0.65, compared to the recall value of 0.22 achieved without normalization for the GC-content bias and copy number. Simply removing normalization for GC-content bias had less of an effect on the total performance; yet adding it had a considerable improvement on prediction accuracy.

We also investigated the effect of using replicates with HMCan-diff on the quality of predictions. We combined all replicates from the second simulation and ran HMCan-diff on the combined data. HMCan-diff efficiently used information from replicates to produce more accurate predictions when compared with combined data predictions (Figure 3A). We speculate that this difference is due to the fact that the variable noise level in different replicates may interfere with the real ChIP-seq signal: it may weaken the ChIP-seq signal in the combined replicates data (Figure 3B).

In theory, to detect differential regions from ChIP-seq data, one could apply a naive two-step approach: first, call peaks for each condition, then, by comparing the presence of peaks at each bin in each condition, define differential regions as regions that have peaks only in one condition. We applied this strategy on the data from simulation 2. We used HMCan (copy number-aware peak caller), and found that HMCan-diff could identify regions with changing density even when HMCan assigned a 'peak' state in both conditions with similar confidence levels (Supplementary Figure S5). Such observations emphasize the importance of the development of a copy number-aware differential peak detector rather than using a copy number-aware peak finder only.

Evaluation on experimental data

With the absence of a gold standard to benchmark regions differentially marked by a histone modification, we decided to carry out an indirect validation. This validation is based on previously observed correlations between the presence of certain histone modifications and transcript levels (3). We selected two different histone marks varying in shape and correlating positively or negatively with gene expression: (i) the H3K27me3 mark linked to Polycomb-based gene silencing and (ii) the H3K27ac mark related to gene activation. Differential gene expression was assessed using the generalized fold change (GFC) (for the cancer sample compared to the normal sample) from the RNA-seq data using the GFold method (34). We checked for correlation between the GFC value and the copy number status, and did not observe any significant correlation that may affect subsequent analyses (Supplementary Figure S6).

First, we examined whether HMCan-diff provides a better read density normalization compared to other methods. We compared the correlation between normalized ChIP-seq read counts (or density values) and GFC for HMCan-diff, csaw, DiffBind, MAnorm, and MEDIPS (Supplementary Figure S7). Methods such as ChIPDiff, MACS2, ODIN and RSEG were excluded from this analysis because they do not provide information regarding normalized read counts in their output. HMCan-diff achieved higher correlation values compared to the other tools on both H3K27me3 and H3K27ac marks. This suggests that HMCan-diff uses a good as or better normalization strategy than the other methods.

Furthermore, we assessed the relationship between the presence of differential histone marks called by each method, and changes in gene expression of corresponding genes. We correlated the predicted differential regions with GFC of gene expression between the cancer samples and their matching healthy samples. We achieved this by running all methods using parameters described in Supplementary Table S6. Then, we associated differential peaks to genes by looking at the overlap of the peaks with 1 kb regions around gene transcription start sites. We used RefSeq (release 68) gene annotation (35). We defined a score S as characterizing the changes in gene expression between the two conditions for a given gene set of size M:

$$S = \sum_{i=1}^{M} \gamma_i log2(GFC_i)/M \qquad (13)$$

Here, γ reflects the sign of the correlation between the histone mark and GFC. In the case of H3K27me3, which correlates negatively with gene expression, we set $\gamma = -1$ for differential regions found in the normal lung tissue and control HMEC cell line, and we set $\gamma = 1$ for differential regions found in the cancer A549 and MCF7 cell lines. In the case of H3K27ac, it positively correlates with gene expression, thus we invert the sign of γ. The S score is similar to the DAGE score (20) and is used to correlate differential regions with gene expression. The higher the value of S, the better the gene expression changes reproduce (predicted) changes in histone modification.

We investigated changes in gene expression corresponding to the top predicted differential peaks for all methods in several scenarios: A549 *vs.* normal lung tissue and MCF7 *vs.* HMEC for histone marks H3K27me3 (Figure 4A and B) and H3K27ac (Supplementary Figure S8A and B). In both comparisons for both marks, HMCan-diff predictions had higher S values (and thus a higher expression fold change) for genes intersecting with differential regions for H3K27me3 (Figure 4A and B) and H3K27ac (Supplementary Figure S8A and Supplementary Figure S8B), compared to other methods.

Moreover, we compared the values of S for the top differential peaks in regions that corresponded to different copy number states: gain, loss and neutral. We obtained those regions by applying Control-FREEC on input samples for the A549 and MCF7 cell lines. In the case of A549 *vs.* lung tissue, HMCan-diff provided noticeably better S values than the other methods in the neutral and gained regions for H3K27me3 (Figure 4C) and in regions of gain and loss for H3K27ac (Supplementary Figure S8C). HMCan-diff also obtained slightly better values in regions of loss for H3K27me3 (Figure 4C) and comparable values in neutral regions for H3K27ac (Supplementary Figure S8C). In the case of MCF7 versus HMEC, HMCan-diff had better S values across all different copy number regions for the H3K27me3 mark (Figure 4D). For H3K27ac (Supplementary Figure S8C), HMCan-diff achieved a noticeably better performance in regions of loss and neutral copy number, and slightly better performance in regions of gain.

Overall, we saw a better performance of HMCan-diff for all copy number regions in all experiments compared to the other methods tested.

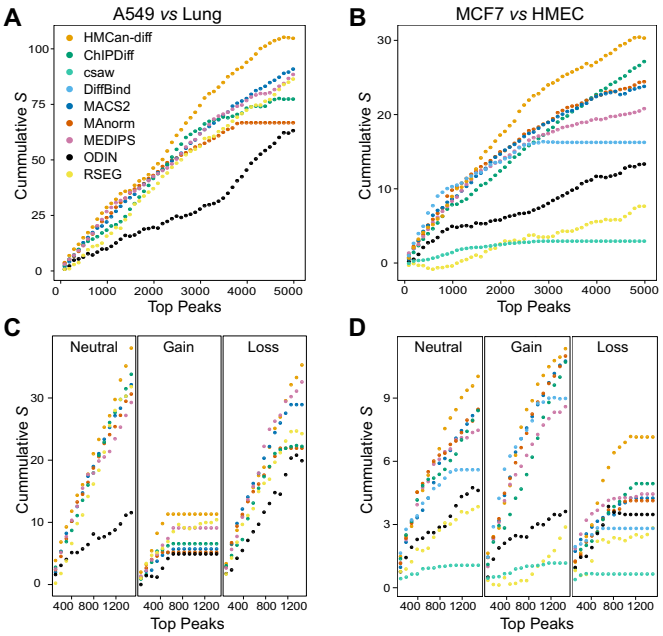

Figure 4. Regardless of copy number status, gene expression changes correlate better with HMCan-diff predictions than with predictions generated by the other methods using the H3K27me3 histone mark. Cumulative values of S corresponding to the top differential peaks when comparing A549 *vs.* normal lung tissue (**A**), and MCF7 *vs.* HMEC (**B**). Cumulative values of S grouped by copy number state (neutral, gain and loss): A549 versus normal lung tissue (**C**), and MCF7 *vs.* HMEC (**D**).

DISCUSSION

We have developed HMCan-diff, a robust method for identifying differences in chromatin modification from cancer ChIP-seq data. HMCan-diff takes into account many covariates that may affect the process of identifying epigenetic changes. The core of what differentiates it from other tools is that in addition to corrections for sequencing depth, it accounts for copy number variation, GC-content bias and variable noise levels. HMCan also utilizes information from replicates if available. Through the combination of these, HMCan-diff identifies differential regions with higher accuracy.

We compared HMCan-diff with eight other tools (ChIPDiff, csaw, DiffBind, MACS2, MEDIPS, MAnorm, ODIN and RSEG). The major conceptual advance for HMCan-diff over these is its method for explicitly correcting for copy number variation. Even though MEDIPS and DiffBind may account for possible copy number alterations (i.e., DiffBind subtracts input to reduce the copy number effect while MEDIPS includes copy number at the final filtering step), our comparison on both simulated and experimental datasets showed that HMCan-diff gave in most cases a dramatically superior performance. The remaining methods do not include any component to handle variable copy number, and are not able to accurately detect differential regions when copy number variation is present.

We noticed some decrease in performance of HMCan-diff in simulation 2 compared to simulation 1 (corresponding, respectively, to the presence and absence of genomic rearrangements between the two conditions). This decrease

was due to a series of normalizations applied by HMCan-diff. It applied some density values in regions with relatively weak simulated ChIP-seq signals close to or below the differential threshold.

HMCan-diff accounts for dispersion present in ChIP-seq data by initializing HMMs based on local Poisson distributions or negative binomial distributions, and defining the emission probability for HMMs as the joint empirical distribution over density values. Also, HMCan-diff assumes that each condition has a single copy number profile, so when using HMCan-diff with replicates, only one copy number profile is estimated per condition.

Similar to other differential peak detection methods, HMCan-diff is designed to work with homogenous samples only. In future, we plan to extend HMCan-diff functionality to handle a mixture of cancer and normal cells. This will enable us to extend the analysis from cancer cell lines to primary tumors.

Although here we provide results of HMCan-diff on histone modification data only, the method is generic and can be applied to other chromatin assays such as DNase-seq (36) and ATAC-seq (37). Conceptually, the use of HMCan-diff is not limited to cancer or even mammalian data. Our method can be applied to compare histone marks between any closely related species. In this case, reads from the two libraries should be mapped to the same reference genome. After analysis, translation of the coordinates of differential regions can be performed using the UCSC liftOver tool.

CONCLUSIONS

We have presented HMCan-diff, a novel computational method that aims to detect differences in histone mark profiles between samples with significant genomic discrepancies. As the principal application of HMCan-diff, we target the comparison of epigenetic profiles between cancer and normal tissue, or between two cancer samples. Associations between genetic events, e.g., mutations in chromatin remodeling genes, and changes in histone modification profiles in cancer, can now be addressed using our method. The method is implemented in C++ and is available at http://www.cbrc.kaust.edu.sa/hmcan.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Sarah Salhi and Kevin Bleakley for English proofreading. The computational analysis for this study was performed on Dragon and SnapDragon compute clusters of the Computational Bioscience Research Center at King Abdullah University of Science and Technology.

FUNDING

KAUST Base Research Funds (to V.B.B. and H.A.); French program 'Investissement d'Avenir', action bioinformatique (ABS4NGS project) (to V.B.); ATIP-Avenir program. Funding for open access charge: ATIP-Avenir program.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science*, **316**, 1497–1502.
2. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
3. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
4. Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
5. Berdasco,M., Ropero,S., Setien,F., Fraga,M.F., Lapunzina,P., Losson,R., Alaminos,M., Cheung,N.K., Rahman,N. and Esteller,M. (2009) Epigenetic inactivation of the Sotos overgrowth syndrome gene histone methyltransferase NSD1 in human neuroblastoma and glioma. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 21830–21835.
6. Ke,X.S., Qu,Y., Rostad,K., Li,W.C., Lin,B., Halvorsen,O.J., Haukaas,S.A., Jonassen,I., Petersen,K., Goldfinger,N. *et al.* (2009) Genome-wide profiling of histone h3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis. *PLoS One*, **4**, e4687.
7. Vendetti,F.P. and Rudin,C.M. (2013) Epigenetic therapy in non-small-cell lung cancer: targeting DNA methyltransferases and histone deacetylases. *Expert Opin. Biol. Ther.*, **13**, 1273–1285.
8. Biancotto,C., Frige,G. and Minucci,S. (2010) Histone modification therapy of cancer. *Adv. Genet.*, **70**, 341–386.
9. Robinson,M.D., Strbenac,D., Stirzaker,C., Statham,A.L., Song,J., Speed,T.P. and Clark,S.J. (2012) Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.*, **22**, 2489–2496.
10. Ashoor,H., Herault,A., Kamoun,A., Radvanyi,F., Bajic,V.B., Barillot,E. and Boeva,V. (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, **29**, 2979–2986.
11. Rashid,N.U., Giresi,P.G., Ibrahim,J.G., Sun,W. and Lieb,J.D. (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.
12. Pickrell,J.K., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, **27**, 2144–2146.
13. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.
14. Chen,L., Wang,C., Qin,Z.S. and Wu,H. (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*, **31**, 1889–1896.
15. Liang,K. and Keles,S. (2012) Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, **28**, 121–122.
16. Xu,H., Wei,C.L., Lin,F. and Sung,W.K. (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.
17. Nair,N.U., Sahu,A.D., Bucher,P. and Moret,B.M. (2012) ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. *PLoS One*, **7**, e39573.
18. Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.
19. Shao,Z., Zhang,Y., Yuan,G.C., Orkin,S.H. and Waxman,D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.
20. Allhoff,M., Sere,K., Chauvistre,H., Lin,Q., Zenke,M. and Costa,I.G. (2014) Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics*, **30**, 3467–3475.
21. Lienhard,M., Grimm,C., Morkel,M., Herwig,R. and Chavez,L. (2014) MEDIPS: genome-wide differential coverage analysis of

sequencing data derived from DNA enrichment experiments. *Bioinformatics*, **30**, 284–286.
22. Zhang,Y., Lin,Y.H., Johnson,T.D., Rozek,L.S. and Sartor,M.A. (2014) PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, **30**, 2568–2575.
23. Lun,A.T. and Smyth,G.K. (2016) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.*, **44**, e45.
24. Fejes,A.P., Robertson,G., Bilenky,M., Varhol,R., Bainbridge,M. and Jones,S.J. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.
25. Boeva,V., Popova,T., Bleakley,K., Chiche,P., Cappo,J., Schleiermacher,G., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.
26. Boeva,V., Zinovyev,A., Bleakley,K., Vert,J.P., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.
27. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
28. Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.
29. Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
30. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
31. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
32. Davis,J. and Goadrich,M. (2006) *Proceedings of the 23rd International Conference on Machine Learning*. ACM, pp. 233–240.
33. Bajic,V.B. (2000) Comparing the success of different prediction software in sequence analysis: a review. *Brief. Bioinform.*, **1**, 214–228.
34. Feng,J., Meyer,C.A., Wang,Q., Liu,J.S., Shirley Liu,X. and Zhang,Y. (2012) GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, **28**, 2782–2788.
35. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
36. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
37. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.